

FMH606 Master's Thesis 2024

IIA-IM

Data-driven Approaches for Pump Condition Monitoring and Curve Estimation

Kristian Sande Sjølyst

Faculty of Technology, Natural sciences and Maritime Sciences
Campus Porsgrunn

Course: FMH606 Master's Thesis, 2024

Title: Data-driven Approaches for Pump Condition Monitoring and Curves Estimation

Number of pages: 65

Keywords: Pump curves, Data-driven models, performance

Student: Kristian Sande Sjølyst

Supervisor: Ru Yan, Saba Mylvaganam

External partner: Martin Forsberg Lie, Dag Harald Skjeltnor

Summary:

The H-Q curve serves as a critical parameter in pump operation and design selection for a given system. However, it is important to recognize that changes over time, such as wear and tear on the pump, pipes, and fittings, as well as alterations to the process and external factors, can significantly impact the pump's performance.

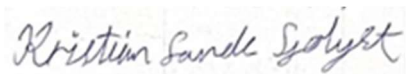
Monitoring changes in the head estimate over time can serve as an early indicator of potential issues, allowing operators or maintenance personnel to take corrective action promptly. The assumption here is that the machine learning models generated can be seamlessly integrated into a control system and executed therein, following a reduction to weights and biases matrices. However, it is worth noting that similar integration can also be achieved through first principles modeling or soft sensors.

This thesis utilizes machine learning, specifically simple neural networks, to achieve positive results in analyzing pump systems. By employing these techniques, the thesis successfully identifies H-Q curves within noisy data, laying the foundations upon which condition monitoring systems can be built. These findings highlight the potential of machine learning in enhancing the understanding and management of pump systems, offering opportunities for improved efficiency, reliability, and predictive maintenance strategies.

Preface

This master's thesis is written in collaboration between the University of south-east Norway and Borregaard. I (the student) have worked at Borregaard for 3 years as a part of the industry master's program has learned a lot from this, and this thesis is the culmination of that experience and new studies to attempt to solve a problem faced in the industry.

Sarpsborg, 15 april 2024

A handwritten signature in cursive script that reads "Kristian Sande Sjølyst". The signature is written in black ink on a light-colored background.

Kristian Sande Sjølyst

Contents

Preface	i
Contents.....	ii
List of tables	iv
List of Figures	v
Nomenclature	vii
List of symbols.....	viii
1 Introduction	1
1.1 Commonly occurring problems in pumps	2
1.2 Interpretation and usage of pump curves	2
2 Existing work and basics	4
2.1 Literature study.....	4
2.2 Standardizing unit for data analysis	6
2.2.1 Converting pressure to meter pump head.	6
2.2.2 Converting Flow measurements.....	7
2.2.3 Calculating Power used by three phase asynchronous motor.	7
2.3 Statistics and t test	7
2.3.1 Mean and standard deviation.....	7
2.3.2 Correlation	8
2.3.3 T test.....	8
3 Test rig setup and Data collection methodology	9
3.1 The test rig and P&ID.....	9
3.2 Experimental design.....	10
3.3 Data Collecting.....	12
3.4 Programing.....	13
3.5 Data pre-processing	13
3.5.1 Artifacts and other junk in the data.....	13
3.5.2 Outliers in the data.....	13
3.5.3 Data Standardization or normalization of the data	14
4 Analytical and data driven methods for H-Q estimates	15
4.1 Estimating a H-Q curve from pump specifications	15
4.2 Recreating pump curve from data	16
4.3 Comparing and analyzing data.....	16
4.4 Preface about sensor packs	17
4.5 Random forest regressor.....	17
4.6 Neural network	17
4.6.1 Activation function.....	18
4.7 Neural network with alternative sensors and different batch sizes	19
4.8 Complementary tests for better insight.....	20
4.8.1 Tanh versus sigmoid activation function	21
4.8.2 More data in the test dataset.....	21
4.8.3 creating two new single output networks for head and watt	21
4.8.4 Adding additional calculated features as inputs	23
4.9 Reducing the network down to its basic weights and biases matrices	23
4.10 Transfer learning from p1001 to p1002	24
5 Data analysis and statistics	26

5.1 Data gathered and pre-processed.....26

5.2 Description of the Data.....26

5.3 Data statistics.....30

5.4 Student t-test.....32

5.5 Data correlation.....33

5.6 Analytical H-Q curve.....37

5.7 Data driven H-Q curve from data gathered37

 5.7.1 *Data driven H-Q curve for p1001*38

 5.7.2 *Data driven H-Q curve for P1002*39

 5.7.3 *Parameter values for both pumps*39

6 Machine learning results40

 6.1 Random forest regressor head estimate.....40

 6.2 Neural network head estimate.....41

 6.2.1 *NN for P1001*.....41

 6.2.2 *NN for P1002*.....44

 6.3 Neural network head estimates using alternative sensors and batch size46

 6.3.1 *NN alternative sensors P1001*.....46

 6.3.2 *NN alternate sensors P1002*.....49

 6.4 Additional tests52

 6.4.1 *Sigmoid versus tanh activation function*.....52

 6.4.2 *Including more data in the test sett*53

 6.4.3 *Single output network results*.....54

 6.4.4 *Adding features*54

 6.5 Reducing the network down to matrix equations55

 6.6 Transfer learning.....55

 6.6.1 *Transfer learning test 1*55

 6.6.2 *Transfer learning test 2*59

7 Discussion.....62

 7.1 Interpolation versus extrapolation.....62

 7.2 Random forest regressor versus neural networks62

 7.3 Correct versus alternative sensors in machine learning.....62

 7.4 Usefulness in condition monitoring63

 7.4.1 *Rule based monitoring*63

 7.4.2 *Smart monitoring*64

 7.4.3 *Examples of rule-based monitoring implementation*64

8 Conclusion66

 8.1 Future work66

References.....a

Appendices.....C

List of tables

Table 3-1 Experimental design table for generating data used to estimate pump curves.....	12
Table 5-1 description of all variables and constants in the dataset.....	27
Table 5-2 statistics for each of the variables in the full dataset showing mean value, standard deviation and unit.....	30
Table 5-3 P1001 t test results showing t statistic, p value for each of the variables level flow, amps, head and watt.....	32
Table 5-4 P1002 t test results showing t statistic, p value for each of the variables level flow, amps, head and watt.....	32
Table 5-5 Correlation for the training data for p1001.....	35
Table 5-6 Correlation for the training data for p1002.....	36
Table 5-7 <i>Hmax</i> and <i>bp</i> estimated values for p1001 and p1002 from the curve fitting process.....	39
Table 6-1 MSE and r^2 scores from the random forest regressor for P1001 and P1002	40
Table 6-2 r^2 scores for p1001 at batch size 16, test score. Over select epochs up to 100	41
Table 6-3 r^2 scores for p1002 at batch size 16 Over select epochs up to 100	44
Table 6-4 r^2 scores of different training method around p1001, test score. Over select epochs up to 100	47
Table 6-5 r^2 scores with different training methods around p1002, test score. Over select epochs up to 100	49
Table 6-6 Test r^2 scores for tanh versus sigmoid activation function. Over select epochs up to 100.....	53
Table 6-7 changed test set for p1001 alternate sensor neural network scores. Over select epochs up to 100	53
Table 6-8 p1001 splitting the network in 2 r^2 scores for alternative sensor measurements batch size 16. Over select epochs up to 100	54
Table 6-9 R^2 test scores for the network with one additional feature. Over select epochs up to 100.....	54
Table 6-10 weights and bias matrices for the alternative sensors p1001 after 100 epochs, shown layer by layer	55
Table 6-11 weights and bias matrices for the alternative sensors p1002 after 100 epochs, shown layer by layer	55
Table 6-12 R^2 test scores for the different transfer learning tests over 25 epochs. Over select epochs up to 25	61

List of Figures

Figure 1-1 H-Q curve, Y-axis is head, X-axis is capacity in L/Min. example of how a manufacturer uses curves as part of marketing material. [1]	3
Figure 3-1 P&ID of the test rig, showing pressure, level, flow indicators, valves and tanks..	10
Figure 4-1 Neural network 1 for pump head and watt estimation given the same inputs as the established models.	18
Figure 4-2 Example Sigmoid curves with $a=-14$, b on label and $c= 14$ from eq (4-9).....	19
Figure 4-3 Neural network with alternative sensors as inputs with watt and head estimate as outputs.....	20
Figure 4-4 One output neural network with alternative sensors as inputs and head as output.	22
Figure 4-5 One output neural network with alternative sensors as inputs and watt as output.	22
Figure 4-6 Neural network with alternative sensors and with one additional feature (flow/level) added. Outputs head and watt.	23
Figure 4-7 Existing alternate sensor neural network with an additional 2 nodes to be trained in transfer learning.	25
Figure 4-8 Existing alternate sensor neural network with an additional two layers to be trained in transfer learning.	25
Figure 5-1 Correlation matrix for the full dataset.....	33
Figure 5-2 Correlation matrix for the training data for p1001. Alternative sensors suit with level, flow, calculated head, pumps amp drawn and calculated watt shown as tags.....	34
Figure 5-3 Correlation for the training data for p1002. Alternative sensors suit with level, flow, pumps amp drawn, calculated head, and calculated watt shown as tags.....	36
Figure 5-4 Estimated HQ Curve made from hand calculating bp and known $Hmax$ on a 10 to 50 L/min flow rate x axis.	37
Figure 5-5 P1001 analytical in green, curve fitted curve in solid blue and error curve in rad with data as blue triangles.....	38
Figure 5-6 P1002 analytical in green, curve fitted curve in solid blue and error curve in rad with data as blue triangles.....	39
Figure 6-1 p1001 training and test loss graph with r^2 scores for p1001 normal sensor dataset.	42
Figure 6-2 Original test data and predictions compared to analytical curve for P1001 normal sensors.....	43
Figure 6-3 Original test data and predictions closer look for P1001 normal sensors.....	43
Figure 6-4 p1002 training and test loss graph with r^2 scores for p1002 normal sensor dataset.	44
Figure 6-5 Original test data and predictions compared to analytical curve for P1002 normal sensors.....	45

List of Figures

Figure 6-6 Original test data and predictions closer look for P1002 normal sensors.	46
Figure 6-7 p1001 alternative sensors batch size 16 training and test loss graph with r^2 scores	47
Figure 6-8 Original test data and predictions compared to analytical curve for P1001 alternative sensors.....	48
Figure 6-9 Original test data and predictions closer look for P1001 alternative sensors	48
Figure 6-10 p1002 alternative sensors batch size 16 training and testing loss graph with r^2 scores.....	50
Figure 6-11 Original test data and predictions compared to analytical curve for P1002 alternative sensors.....	51
Figure 6-12 Original test data and predictions closer look for P1002 alternative sensors.	51
Figure 6-13 p1001 alternative training data with tanh activation function.....	52
Figure 6-14 original model trained on data from p1001 with the alternate sensors.	56
Figure 6-15 transfer learning model trained on data from p1002 using the original model from p1001, alternative sensor set.....	57
Figure 6-16 Original test data and predictions compared to analytical curve for P1002 alternative sensors using transfer learning model 1.....	58
Figure 6-17 Original test data and predictions closer look for P1002 alternative sensors using transfer learning model 1.	58
Figure 6-18 transfer learning model with 2 additional layers with a sigmoid activation function.	59
Figure 6-19 Original test data and predictions closer look for P1002 alternative sensors using transfer learning model 2.	60
Figure 6-20 Original test data and predictions closer look for P1002 alternative sensors using transfer learning model 2.	60
Figure 7-1 p1001 data plot with high and low limits in orange at 30 and 20 L/min respectively.	65
Figure 7-2 p1002 data plot with high and low limits in orange at 30 and 20 L/min respectively.	65

Nomenclature

CSV – comma separated values.

Daca – data acquisition

H-Q – Pump Head [m] over Volume flow rate [L/min]

Inf – infinity

ML – machine learning

NaN – not a number

NN – neural network

NPSH – Net positive suction head [m]

NPSHr – Net Positive Suction Head requirement [m]

Pida- PID controller

PID - proportional integral derivative

P&ID – piping and instrumentation diagram

RFR – random forest regressor

SP – shaft power [kW]

VFD – Variable frequency drive

List of symbols

p	Pressure, [<i>pascal</i>]
g	Gravitational constant, [$\frac{m}{s^2}$]
h	Height, [m]
H	Head, [m]
\dot{m}	Mass flow rate
\dot{V}	Volume flow rate, [m^3/h]
ρ	Density, [$\frac{kg}{m^3}$]
P	Power, [<i>Watt</i>]
U	Volt, [V]
I	Ampere, [A]
φ	Phase change angle
a, b, c	Scaling constants
Q	Flow rate, [$\frac{L}{min}$]
z_n	Network layer n
W_n	Weights matrix n
b_n	Bias matrix n
a_n	Weights matrix post activation function n
\hat{y}	Output from neural network
x	Input vector to network

1 Introduction

The processing industry extensively utilizes pumps, incurring significant operational and maintenance costs. So, if there exists a method to know or have a good estimate of the pump's performance based on what the operators already know, then there should also be a way to perform preventative maintenance or operate the pumps closer to an optimal point to increase its lifespan, without affecting the overall plant performance. But what does the operator know about how the pumps perform? The answer to that is shockingly limited as the operator normally only has an idea of the speed of the pump and the output from the pump is low. There is not common to have the sensors; pressure in, pressure out and flow out of every pump in a plant and not even a handful will have all the necessary sensors to give the operators a good idea where the pump is on its relevant curves at any given time.

This thesis will attempt to answer the question of pump performance by the method of data analysis and machine learning. More specifically, the work will investigate how data driven models can be applied to detect failure and optimize operation and maintenance schedules. By leveraging those tools look at pump head and pump curves from familiar and unfamiliar pump systems. With the end goal of giving the operators and maintenance team a better understanding of how the pumps is being operated historically and their current condition.

This thesis will start off with two pumps in a controlled environment, outfitted with some relevant sensors installed. Installed sensors include pressure inn, pressure out, flow out, levels of tanks, speed setting, and amps drawn by the pump. Subsequent sections will discuss these sensors and the test rig in more detail.

The thesis structure comprises of an introductory section providing an overview of the basic content and concepts utilized in greater detail later. Following this a chapter dedicated to previously established theory including a literature study and some relevant mathematics that will be used in preprocessing and analysis later.

Subsequently five chapters detailing the methods used for the test rig and experimental design, data analysis from the test rig, two methods of generating H-Q curves, diverse ways to utilize machine learning for all of this and last some notes on condition monitoring.

The methods used and results obtained are presented in each chapter based on the tasks defined. Relevant data, information on codes and test procedures are presented under the section "Appendix".

1.1 Commonly occurring problems in pumps

Several prevalent problems encountered by pumps in operational environments include, but are not limited to, the following, listed without any specific order: [4]

- Faulty seals
 - The fittings and housing for the pumps degraded over time with faulty operating conditions.
- Cavitation
 - Air and other things cause damage to the internal housing of the pump, reducing performance.
 - Usually caused by insufficient pressure at suction port.
- Operation off breakeven point
 - Operating “off the curve”, either too far left or right of the pumps designed operating point.
 - This will be the main research question of this thesis.
- Excess friction
 - Lack of lubrication.
 - Excess vibration.
 - Caused by a lack of preventative maintenance schedule or insufficient lubrication during scheduled maintenance.
- Faulty ball bearings
 - General wear and tear causing the bearings to fail over time.
 - Requires vibration monitoring to detect, which was not installed on the test rig, hence not included in this thesis.
- Pipes clogged.
 - No more flow

Certain faults may not be discernible from available data or may only become apparent upon pump failure. Consequently, conducting a study on the remaining useful life within the timeframe of this report is unfeasible.

1.2 Interpretation and usage of pump curves

Figure 1-1 illustrates the potential approach a company might undertake to advertise their pumps and curves, presenting a wealth of information that may pose readability challenges to individuals lacking prior experience with pump curves. This summary encapsulates the content of the book in ref [3] without delving into extraneous specifics. Figure used mostly for illustrative purpose and not as facts.

It is important to note a discrepancy between the flow rates specified for the pumps used in the test rig, which are rated for 10 to 50 L/min, and the range indicated on the graph, which extends from 10 L/min to 160 L/min. Additionally, the pump head recorded at 10 L/min appears inconsistent with the data provided by the manufacturer, which states a head of 14m at this flow rate. This discrepancy discussed will not lead to any problems in the analysis, as the pumps are known to be different with different characteristics.

The pumps in the test rig are MX-250 with impeller mark 5 [21], all lines referring to something else relevant.

The line labeled NPSHr (Net Positive Suction Head required) at the bottom of the graph denotes the minimum necessary pressure at the suction port for the pump to function properly. Decreasing NPSH can lead to cavitation and subsequently damage to the internals of the pump.

Above the NPSHr line, the H-Q curve illustrates the relationship between the pump's head pressure and its flow rate. This curve indicates that as the flow rate decreases, the pump can generate higher pressure, and conversely, as the flow rate increases, the pressure decreases. It is crucial to note that this flow is contingent upon the resistance encountered by the pump rather than the speed at which the pump operates. Furthermore, altering the speed of the pump will inevitably affect the pump curves. With a lower operating speed, the H-Q curve itself will decrease relative to the speed setting of the pump. The max head it can generate will be lower and max flow rate will be lowered as well.

The final curve to consider is the shaft power curve, which represents the power requirement necessary for the pump to sustain the existing flow rate. Shaft power and motor power are not the same, shaft power will always be lower than motor power by a known effectiveness factor.

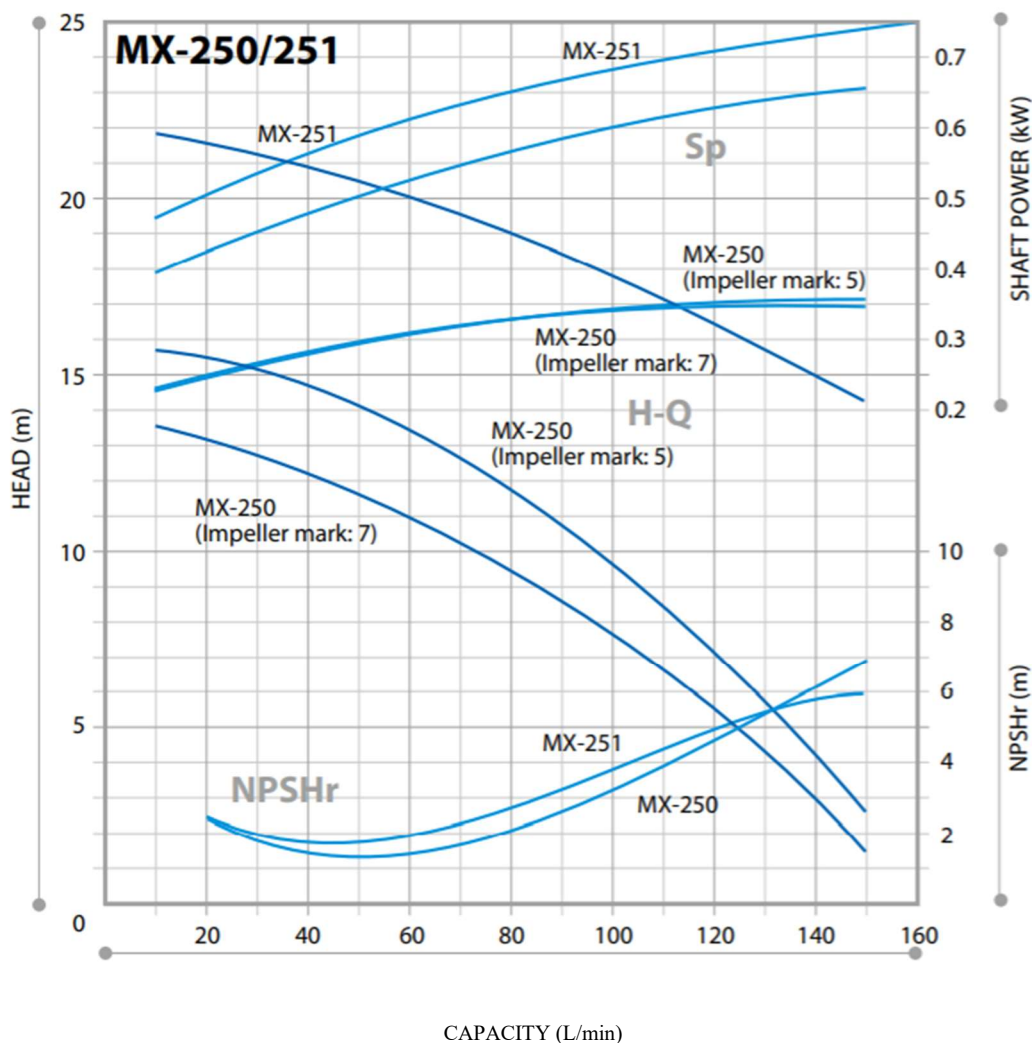


Figure 1-1 H-Q curve, Y-axis is head, X-axis is capacity in L/Min. example of how a manufacturer uses curves as part of marketing material. [1]

2 Existing work and basics

This chapter will encompass established theoretical frameworks that will be subsequently applied in later stages of the study. Including the literature study, how to do basic unit conversions and some statistics.

2.1 Literature study

Considerable literature regarding pump performance monitoring. Among the most influential sources were “Understanding pump curves” [3], which provided the fundamental insight into the thesis objectives, including the utilization of available sensor information and methodology for calculating pump head. Although the referenced book primarily employed imperial units, those units were easily converted to fit the equations later.

Drawing upon the insight gained from both “Optimization Methodology for Estimating Pump Curves Using SCADA Data” [2] a conceptual framework for data utilization and curve fitting began to take shape. This conceptualization was inspired by the methodology outlined in these sources, facilitating the gathering of pertinent data, and guiding the application of curve fitting techniques to reconstruct an approximation of the pump system's specific pump curve. Furthermore, valuable contributions to my understanding of this topic were provided by my colleague, Carsten, a mechanical engineer immersed in daily interactions with pump systems [4].

At this point in the study, while the methodologies for constructing a pump curve from data were established, the integration for machine learning techniques had not yet crystallized. It was not until reading parts of the article “Train longer, generalize better: closing the generalization gap in large batch training of neural networks” [5] that a realization was made. This article prompted the recognition that the requisite X and y data for designing a neural network was already at hand. Here, X denotes the input data matrix and y represents the output vector.

The influence of [5] extends beyond its own content, it reinforced the principles in “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems” [9] . This text reignited my focus on implementing ML with Keras and TensorFlow. It comprehensively covered network setup and training methodologies. However, recent updates to TensorFlow necessitated a shift to PyTorch for network configuration, utilizing its online documentation [10]. While the transition introduced a change in platform, the fundamental concepts such as training duration and batch size persisted. Drawing from PyTorch resources, I explored various activation functions, selecting sigmoid and tanh for their suitability in capturing the nonlinear characteristics of pump curves.

Additionally, marketing materials [1] were consulted to gain insights into how pump manufacturers articulate information about their products. These resources served as a benchmark for envisioning the potential outcome of the study's efforts and facilitated the elucidation of the study's fundamental objectives.

Earlier in this subchapter, the concept of X and y matrices was introduced, raising the question of how data is incorporated into these matrices. Drawing from insight gained from “Introduction to Engineering Experimentation” [6][8] a robust framework was established for

2 Existing work and basics

designing a series of experiments to be conducted on the test rig provided by Borregaard. This framework prompted consideration of the type and volume of data required. Armed with an understanding of how the pump behaves within its operational parameters and equipped with data from the test rig's measurements, the control system was configured to manipulate valves, enabling data collection for the X data matrix.

Subsequently, the data from X underwent preprocessing and computation to determine the corresponding y values, thereby completing the full dataset. This dataset is deemed comprehensive for generating pump curves based on both collected data and pump specifications. Furthermore, it was partitioned into training and testing datasets to facilitate machine learning processes. While the endeavor resembles the development of a soft sensor, it remains distinct in that it operates offline, maintaining its place solely within the dataset.

Typically, when assessing the condition of a pump, both its position on the curve concerning head and flow, as well as the power rating of the motor, are crucial factors to consider. However, determining the wattage consumed by the motor was a task I had prior experience with. To refresh my understanding of three-phase power systems and the principles governing phase changes over coils, I turned to the book “Fundamentals of electric power engineering: From electromagnetics to power systems” [7]

Several additional works were initially considered for inclusion in the study; however, they were eventually excluded due to their methodologies falling outside the scope of this thesis. For instance, “Condition monitoring for early failure detection. Frognerparken pumping station as case study” [11] presents a case study aiming to develop a condition score ranging from 100 to 0 for pumps using temperature and amperage readings from each coil, employing a Kalman filter. While this approach demonstrates promise, the necessary measurements were not accessible for the pumps involved in this study. Additionally, vibration analysis was intended to be incorporated, but due to delays, the required sensors did not arrive in time.

Regrettably, one of the most intriguing papers that had to be excluded was “Deep learning for centrifugal pump condition monitoring using data from variable frequency drive” [12] This study focused on detecting cavitation in pumps by monitoring the NPSH (Net Positive Suction Head), with reductions of up to 3% indicating cavitation. However, its implementation required a valve upstream of the pump to regulate flow and pressure, which was not available in the test rig used for this study. Although a manual valve existed between the tank and the pump in the test rig, it was deemed too imprecise for conducting experiments.

While this thesis will primarily leverage machine learning techniques, it's important to acknowledge the value of both soft sensors and first principles modeling in the context of pump condition monitoring. This study opts for machine learning as the soft sensor approach, thereby minimizing the necessity for delving into complex mathematical models. The assumption here is that the machine learning models generated can be seamlessly integrated into a control system and executed therein, following a reduction to weights and biases matrices. However, it's worth noting that similar integration can also be achieved through first principles modeling. Ultimately, the choice between methods is driven more by practical effectiveness than a strict determination of superiority.

The paper “Estimation of neurons and forward propagation in neural net” [20] delves into the estimation of neuron weights manually, offering insights into understanding neural networks

not merely as black box models but as sets of equations. This perspective proves valuable when simplifying the network to matrices.

2.2 Standardizing unit for data analysis

To ensure consistency and facilitate a more scientific analysis of the data, it is imperative to standardize units where necessary. Although certain nonstandard or industry-specific units may be easier to comprehend and operate with, converting them to standard units is essential for rigorous scientific investigation. The conversion process is typically straightforward and will be detailed in subsequent subchapters to maintain clarity and precision in the analysis.

2.2.1 Converting pressure to meter pump head.

To convert a pressure reading from bar to meter a few assumptions are necessary to be made:

- The density of the media is known and uniform.
 - o Most materials change based on temperature, if not completely solidify or vaporize.
 - o The media can change from day to day so it will not be constant but in a known range.
- Physics constant g can still be 9.81 and not needed to recalibrate it for where we are.
- The fluid is incompressible.
- Friction losses are neglectable.
- Hydro static losses are unknown.

With those assumptions and the formula (2-1).

$$\Delta p = \rho g H \quad (2-1)$$

$$\Delta p = p_{inn} - p_{out}$$

Where p is the pressure in pascal, g is the gravitational constant 9.81 and H is the pumps head in meter.

Rewriting the formula to solve for H (2-2). And including the loss terms assumed to be 0.

$$H = \frac{\Delta p}{\rho g} + h_{friction} + h_{static} \quad (2-2)$$

representing H [m] as a function of pressure; $H(\Delta P)$. With $h_{friction}$ being the loss of head due to friction of the fluid against the pipe and h_{static} is the loss of head due to high of the pipes

There will be a need to convert from bar to pascal, but that is also just a linear transformation where $1 \text{ Bar} = 10^5 \text{ pascal}$. The unit of the control system is in mBar, so the following will be used; $1 \text{ mBar} = 10^{-3} \text{ Bar} = 10^2 \text{ pascal}$.

The calculation of head losses resulting from friction and static pressure within the system's configuration is considered beyond the scope of this thesis. Nonetheless, it is crucial to recognize these factors as they play a significant role in comprehending potential deviations from the expected maximum head observed during the experiments. By acknowledging these

factors, the thesis aims to provide a comprehensive overview of the experimental outcomes while recognizing the broader context of hydraulic system dynamics.

2.2.2 Converting Flow measurements

To convert the flow rate from liters per hour (L/h) to liters per minute (L/min), a simple solution is to divide by 60, as there are 60 minutes in an hour. This conversion allows for consistency in units and facilitates easier comparison and analysis of flow rates.

When dealing with flow rates in kg/h, a direct conversion to L/min cannot be performed as additional considerations are necessary due to material density. Assuming the fluid's density is known, equation (2-4) can be utilized to convert from mass flow to volume flow and the same 60 min/h trick still applies.

$$\dot{m} = \dot{V} \rho \quad (2-3)$$

$$\dot{V} = \frac{\dot{m}}{\rho} \quad (2-4)$$

Where \dot{m} is the mass flow of the fluid in kg/min, \dot{V} is the volume flow in m³/min and ρ is the fluid's density in kg/m³.

2.2.3 Calculating Power used by three phase asynchronous motor.

To accurately determine the power drawn from the grid by a 3-phase asynchronous motor, the watt needs to be calculated with knowledge gained from looking at the pump's sign. This is where it will say its effectiveness and phase change. With that knowledge and equation (2-5) the power can be calculated.

$$P = U I \sqrt{3} \cos(\varphi) \quad (2-5)$$

Where P is watt used, U is volt, I in amps, $\sqrt{3}$ is due to 3 phases and $\cos(\varphi)$ is the power lost due to phase shifting. [7]

2.3 Statistics and t test

Understanding the dataset on a deeper level than merely examining numerical values is paramount. This subchapter delves into methodologies aimed at precisely achieving that.

2.3.1 Mean and standard deviation

The mean or average value of a dataset provides valuable insight into the central tendency of a variable in the data, indicating where the majority of observations cluster. When combined with the dataset's standard deviation, which measures the extent of variability or dispersion of the data points, a comprehensive understanding of the dataset's characteristics emerges.

Standard deviation is closely related to variance, offering a measure of how much individual data points deviate from the mean.

2.3.2 Correlation

While mean values and standard deviation looks at a single variable in the data, correlations coefficients quantify the degree of linear relationship between two variables, meaning that as one variable increases the other tends to increase as well. Conversely a common coefficient close to -1 signifies a strong negative correlation indicating that as one variable increases the other tends to decrease. A correlation of 0 suggests no linear relationship between the variables.

2.3.3 T test

Before utilizing the dataset, it is prudent to comprehend the nuances, disparities, and fluctuations it contains. Employing significance testing, notably the Student's t-test, proves invaluable in this endeavor. This statistical method offers insights into the significance and probability of disparities between the test and training datasets, thereby facilitating a more informed analysis.

The null hypothesis, commonly regarded as the default assumption, asserts that there exists no significant difference between the groups or datasets under scrutiny. The T-statistic serves to quantify the relative distinction between the mean values of the two datasets, while the P-value indicates the likelihood of observing the data assuming the null hypothesis is valid. These statistical measures provide valuable insights into the comparability and significance of the datasets being analyzed.

Interpreting the outcomes derived from these statistical metrics holds significant sway over the utilization of the data. Depending on the observed values within the datasets, decisions regarding the suitability and relevance of the data may be interpreted differently if a high degree of similarity is found. These insights inform the researcher's judgment regarding the reliability and applicability of the datasets under examination, thereby guiding subsequent analytical and interpretive endeavors.

methodology

3 Test rig setup and Data collection methodology

3.1 This chapter will describe the test rig setup and the methodological framework employed for data acquisition. This encompasses the construction of the piping and instrumentation diagram (P&ID), along with the experimental design implemented for data collection. Furthermore, this chapter will describe the methodologies employed for data processing, encompassing data acquisition procedures, programming techniques, library utilization, as well as outlier detection and data error mitigation strategies. The objective is to culminate with a curated dataset ready for subsequent analysis and machine learning. The test rig and P&ID

The test rig consists of a few different components as seen in the P&ID, Figure 3-1.

This rig has the following sensors:

- Four (4) pressure sensors. Range 0 to 1.5 bar gauge

These sensors are used to indicate the pressure inn and pressure out of the pumps, in pascal, and are tagged PI1000, PI1011 around P1001 and PI1013A, PI1013B around P1002.

- Two (2) flow sensors. Range 0 to 50 L/min

These flow sensors are used to indicate the rate of flow out of the pumps, FiC1003s process value, PV, after P1001 and Fi1015 after P1002. Flow is indicated in L/min. Both magnetic flow transmitters.

- Two (2) level sensors. Range 0-100% tank level

These sensors work on different principles but will give the same indication of level in %. LI1001 on tank B1001 is a radar device and used time of flight to calculate the tanks level, and LIC1005 PV is a pressure transmitter used to determine the level in tank B1002.

- Two (2) flow control valves. Operating range 0 to 100% opening

The valves are connected in the control system to a PID controller, hence why some of the measurements are the process value from a PID controller. These valves will then be the operation point, OP, of the controller. The first one after P1001 is FIC1003.OP and sits about 4m off the ground into tank B1002. And the second one is LIC1005.OP mounted at the same height as P1002, after P1002.

methodology

While all described sensors will undergo equal preprocessing, the pressure sensors will serve a specific analytical function: calculating the pump head. This process resembles that of a soft sensor, albeit implemented solely within the dataset rather than in real-time online monitoring. It's important to note that these pressure sensors will not be utilized in all machine learning tests.

The level indicators will be employed in select machine learning methodologies instead of the measurements from dedicated pressure sensors, thus offering alternative process measurements.

Flow will be used in head calculation, curve fitting and ML methods later.

While valve operating points will not directly feature in the machine learning algorithms, they will play a crucial role as parameters for constructing training and test datasets. Furthermore, they are pivotal in the data gathering process.

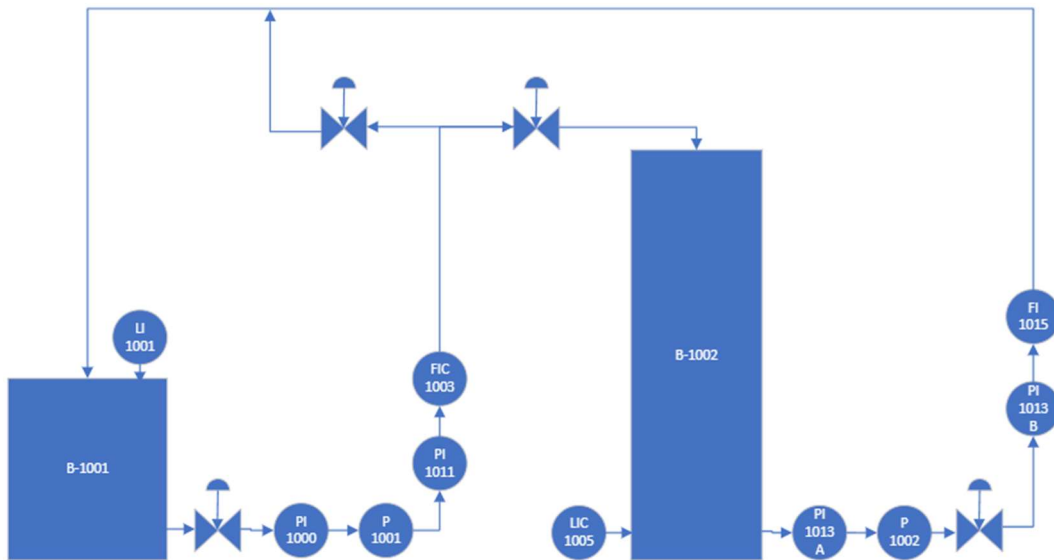


Figure 3-1 P&ID of the test rig, showing pressure, level, flow indicators, valves, and tanks.

3.2 Experimental design

Creating a comprehensive dataset necessitates a meticulous approach to experimental design [6]. To generate a pump curve experimentally, it's imperative to understand the underlying process of curve generation, which involves controlling the valves downstream of the pump. This manipulation induces a constriction or resistance in the flow that the pump must overcome, resulting in an increase in pressure as the flow decreases. The experimental design is anchored on this principle, with the overarching objective of recreating the pump curves through curve fitting in subsequent analyses, as elaborated further in later discussions.

Table 3-1 outlines the valve openings against which the pumps will be operated, along with the corresponding variables to be collected. These variables align with those discussed in Section 3.1, encompassing 2 flow measurements, 2 suction pressure readings, 2 outlet pressure readings, 2 amperage measurements, 2 speed settings, and ambient temperature.

methodology

This data collection process will be repeated for every desired speed setting at which the pump may operate, this thesis will exclusively focus on the pump's nominal speed setting [4]. The 'X' indicates all parameters that will be logged during the experiments.

methodology

Table 3-1 Experimental design table for generating data used to estimate pump curves.

Valve opening	0%	25%	50%	75%	100%
Flow outlets	X	X	X	X	X
Pressure outlet	X	X	X	X	X
Pressure suction	X	X	X	X	X
Amperes	X	X	X	X	X
Pump speed	X	X	X	X	X
Ambient temperature	X	X	X	X	X

Each column of the table will be executed as a distinct experiment, with data collected at 10-minute intervals, repeated multiple times.

It's crucial to acknowledge that operating pumps against a closed valve results in the dissipation of energy within the system, potentially leading to overheating. Consequently, this approach is not advisable.

This experiment will serve to generate the data necessary for curve fitting and training the neural network. However, it's also essential to have a separate test dataset. This dataset will involve sweeping over valve openings not included in the original experiments, such as 80-95%, 55-70%, 30-45%, and 5-20%. While the test dataset does not require the same scale as the training dataset, it will still be run for a few hours within those specified ranges to ensure its effectiveness.

3.3 Data Collecting

In the preceding subchapter, Figure 3-1 illustrated the test rig, which will serve as the primary source of data collection. However, simply configuring the valves correctly and initiating the run will not yield a valuable dataset.

The control system facilitates data gathering by exporting .CSV files with one-second resolution for all activities occurring within the last hour. This method will serve as the primary means of data collection for subsequent analysis in Python or other software platforms.

methodology

3.4 Programing

Code is developed in collaboration with ChatGPT 3.5 [13]. That said, it is not flawless and needs manual fixing most of the time. The following libraries will be used:

1. NumPy
 - a. For minor functions and basic data handling[14]
2. Pandas
 - a. Builds on NumPy and will be used for its data frames and processes related to data frames. [15]
3. SciPy
 - a. Contains methods for optimization problems and curve fitting. [16]
4. Matplotlib
 - a. To plot things. [17]
5. Sklearn
 - a. Data preparation for machine learning.
 - b. Curve fitting methods.
 - c. Random forest regressor. [18]
6. PyTorch
 - a. Used to build, train, and test neural networks and other machine learning algorithms. [10]

This thesis will not heavily focus on in-depth code analysis.

3.5 Data pre-processing

All data must be standardized to the same units and format. Subsequently, pressure will be converted to meter pump head [m], and all flows will be standardized to liters per minute [L/min]. Additionally, there will be a requirement to filter or remove invalid data collected and to normalize or standardize the dataset. These processes will be elaborated upon in subsequent subchapters.

3.5.1 Artifacts and other junk in the data

The raw data may contain characters such as '\xa' within the numbers, which need to be removed to ensure data quality. Additionally, instances of 'inf' (infinity) and 'NaN' (Not a number) are to be avoided in the dataset.

3.5.2 Outliers in the data

It's highly likely that some data points may deviate significantly from the rest of the dataset, presenting outliers that can pose challenges for curve fitting and machine learning algorithms unless appropriately addressed.

Identifying outliers during experiments involves considering various factors. For instance, if the rig is not in operation yet certain trends persist or if the values remain consistent from non-operational conditions, such instances could be considered outliers. These outliers can be filtered out by examining variables such as minimum flow and head, as they are expected to be close to zero under non-operational conditions.

methodology

3.5.3 Data Standardization or normalization of the data

Standardization involves bringing all values to a mean of 0 with the same normal distribution as the original data. This results in every constant value being 0, while the remaining values are normally distributed within the specified range.

On the other hand, normalization scales all values to the range [0,1] based on the minimum and maximum found values in the variable. While normalization is more susceptible to noise processes and significant outliers, it could still work well for datasets like the one gathered in this thesis, which includes values ranging from 0-1200mbar, 0-100%, and 0-1.5A.[8]

estimates

4 Analytical and data driven methods for H-Q estimates

This chapter will employ the dataset previously processed in Chapter 3 to estimate the H-Q curve, incorporating both established specifications and data acquired from the test rig. A comparative analysis will be undertaken between these two methodologies to construct an error curve. Moreover, this chapter will comprehensively cover all facets pertaining to machine learning, encompassing the methodologies utilized and pertinent theoretical deliberations. Multiple tests will be conducted, with each explained in subsequent subchapters.

4.1 Estimating a H-Q curve from pump specifications

Drawing inspiration from Reference [2], the process involves fitting a quadratic curve to the data obtained from the Scada system. This approach is particularly viable in this context, given the pre-existing knowledge of the pumps involved.

$$H_p = H_{p,max} - b_p Q_p^2 \quad (4-1)$$

Utilizing equation (4-1) where H_p represents the head at the given flow, $H_{p,max}$ denotes the pumps maximum head capacity as specified by the manufacturer, b_p is a parameter subject to optimization or calculation, and finally Q_p represents the measured flow discharged by the pump.

In this context, pumps are known, with $H_{p,max} = 14m$ and the flow Q_p falls within the range $10 \leq Q_p \leq 50$. Consequently, H_p must reside within the range $14 \geq H_p \geq 0$, utilizing these constraints, a candidate for b_p can be derived, leading to the formulation of the equations at the limits given by (4-2).

$$14 = 14 - 100b_p \rightarrow b_p = 0 \quad (4-2)$$

$$0 = 14 - 250b_p \rightarrow b_p = \frac{7}{125}$$

Those equations yield different values for b_p rendering them non-analytic in nature.

attempting to offset the flow by the pumps minimum rated flow yields the modified equation (4-3).

$$H_p = H_{p,max} - b_p(Q_p - Q_{p,min})^2 \quad (4-3)$$

$Q_{p,min}$ denotes the minimum rated flow of the pump when it is nonzero, as is the case here.

And using eq (4-3) with the limits discussed earlier in this sub chapter gives, (4-4)

estimates

$$14 = 14 - b_p(10 - 10)^2 \rightarrow 14 = 14 \quad (4-4)$$

$$0 = 14 - b_p(50 - 10)^2 \rightarrow b_p = \frac{7}{800}$$

Now an analytical method for acquiring b_p exists.

All these calculations are predicated on the assumption that the pump will adhere to its specified behavior and operate at nominal speeds.

4.2 Recreating pump curve from data

For an initial examination of a data-driven model aimed at estimating the pump's H-Q curve from available data, the dataset comprises variables detailed in Chapter 3.2, acquired at a resolution of 1 second directly from the source. Following preprocessing outlined in Chapter 3.5, the data is prepared for analysis.

In this instance, while the pumps are identified, future attempts may involve unknown pumps. Hence, the pumps are treated as unknowns. Consequently, at least two parameters remain unknown when estimating a pump curve using equation (4-1): $H_{p,max}$ and b_p .

To establish a methodology for estimating these parameters from data, the first step is to devise a method for minimizing error.

$$\varepsilon_H = H_{est} - H_{obs} \quad (4-5)$$

$$H_{est} = H_{max} - b_p Q_{obs}^2 \quad (4-6)$$

$$H_{max}, b_p > 0 \in \mathbb{R} \quad (4-7)$$

Here H_{obs} represents the calculated head derived from pressure sensors or observed values. H_{est} signifies the head estimated by curve fitting based on available data and observed flow and ε_H denotes the error to be minimized.

Utilizing equation (4-5) as the objective function to minimize, (4-6) as the equality constraints, and (4-7) as the inequality constraints and boundaries for a curve fitting optimization problem for determine \hat{H}_{max}, \hat{b}_p , thereby reconstruct the curve from 0 in Python.

4.3 Comparing and analyzing data

Comparing the results obtained from the analytical approach with those derived from data-driven models is a straightforward yet intricate task. Assuming the data's quality is adequate, one simple measure for comparison can be represented as equation (4-8).

$$\mathcal{E}_T = |H_A - H_D| \quad (4-8)$$

estimates

The total error is defined as the absolute difference between the analytical curve and the data-driven curve. This calculation yields a new curve spanning the flow range, encapsulating the errors at different points. Additionally, this error curve can be plotted alongside the analytical and data-driven curves for visual comparison.

4.4 Preface about sensor packs

The "normal sensors" encompass the essential sensors required for conducting analytical methods to calculate pump head. These sensors typically include pressure in, pressure out, flow out, and the amperes drawn by the pump. Additionally, if a variable frequency drive is utilized, hertz measurements are also necessary.

On the other hand, the "alternative sensors" aim to represent a more practical set of sensors commonly found in process plants. These sensors include the level of the tank in front of the pump, flow out, and power drawn. Similar to the normal sensors, hertz measurements are included if a variable frequency drive is employed.

4.5 Random forest regressor

Utilizing a random forest algorithm involves constructing multiple decision trees and subsequently averaging their outputs. This classic and straightforward approach in machine learning will be tested using the alternative sensors mentioned earlier, which will also be employed in subsequent neural network experiments.

Instead of relying on pressure in, pressure out, and flow out data from the pump, the random forest model will utilize data from the level of the tank in front of the pump, the flow out rate, and the amperes drawn by the pump. With this information, the random forest model will estimate the pump's head. This serves as a foundational point of comparison for subsequent analyses.

4.6 Neural network

The advantage of employing a neural network over traditional machine learning algorithms lies in its ability to represent even complex networks as straightforward matrices of weights and biases. This characteristic facilitates seamless integration into control systems.

To train a simple neural network to model the pump data, the network requires pressure in, pressure out, flow out, and amperes as inputs, while head and amperes serve as outputs, forming the X and Y vectors, respectively.

The neural network architecture consists of 4 input nodes, 5 nodes in the hidden layer, and 2 output nodes. It employs the sigmoid function as its activation function, as depicted in Figure 4-1.

estimates

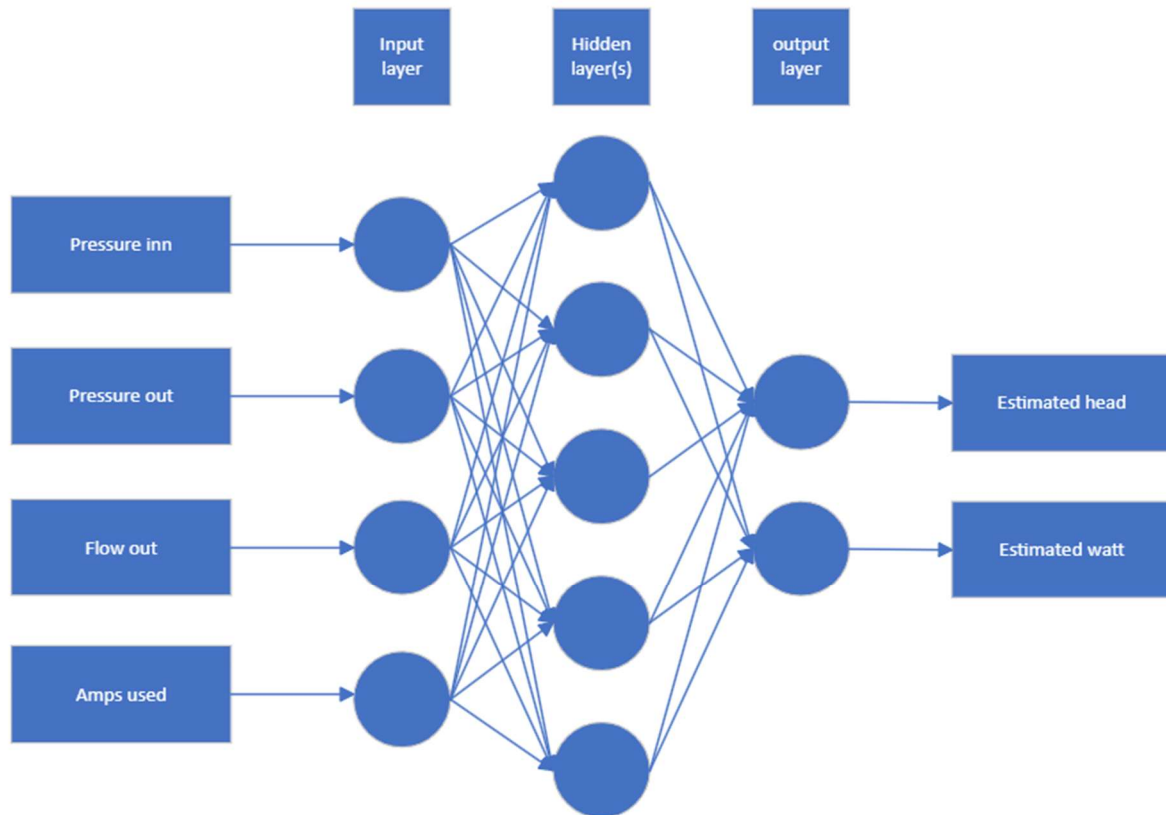


Figure 4-1 Neural network 1 for pump head and watt estimation given the same inputs as the established models.

4.6.1 Activation function

The Sigmoid activation function is selected due to its non-linear nature. This characteristic provides the neural network with an understanding of the type of data it processes. Given that pump curves exhibit polynomial behavior, the Sigmoid function can adapt its curve to fit such data.

Equation (4-9) illustrates a generalized sigmoid function with three parameters; amplitude (a), slope (b), and offset (c). Figure 4-2 demonstrates how the sigmoid curve can be adjusted using these parameters to fit the data. In the figure, the parameters are set as follows: $a = -14$, $b = [1, 0.75, 0.5, 0.25]$, and $c = 14$. These settings cause the curve to range from 14 to 0, as opposed to 0 to -14. It's important to note that this example showcases how a sigmoid curve can be utilized and not necessarily how it functions directly within the neural network.

$$f(x) = \frac{a}{1 + e^{-bx}} + c \quad (4-9)$$

estimates

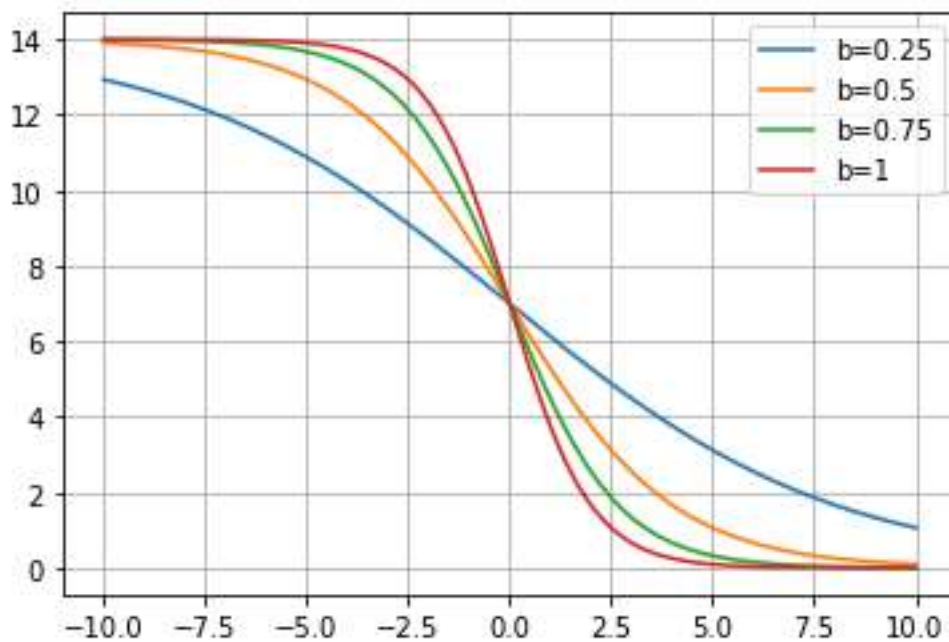


Figure 4-2 Example Sigmoid curves with $a=-14$, b on label and $c=14$ from eq (4-9)

4.7 Neural network with alternative sensors and different batch sizes

In real-world production plants, it's uncommon to have all the ideal measurements (such as pressure in, pressure out, and flow out) for calculating the pump's head. Instead, it's more typical to rely on measurements like the level of a nearby tank and the flow out of the pump, with sparse or no pressure data available.

Given this scenario, machine learning can be leveraged to predict the pump's head using alternative measurements and compare it to the traditional mathematical approach. Figure 4-3 illustrates the neural network architecture designed for this purpose, featuring 3 inputs, 1 hidden layer with 4 nodes, and 2 outputs. The inputs include LI1001, FIC1003, and p1001, with outputs H1 and watt1 for the learning around p1001. A similar architecture is employed for p1002, with inputs LIC1005, FI1015, and p1001. In both cases, the inputs consist of the level of the tank in front of the pump, the flow out of the pump, and the current drawn by the pump.

Activation functions are applied to the layers, with sigmoid used for layer 1 and linear for the remaining layers. The architecture adopts a 3-node to 4-node to 2-node configuration to prevent overfitting or learning extraneous features.

When training a network, two crucial parameters to consider are batch size and training time. The experiment will evaluate the testing accuracy of the model using different batch sizes, ranging from small to large. [5]

estimates

For assessing the accuracy of a regression model, the R^2 score is preferred as it measures the correlation between the output and input data. This metric provides valuable insights into the model's predictive performance.

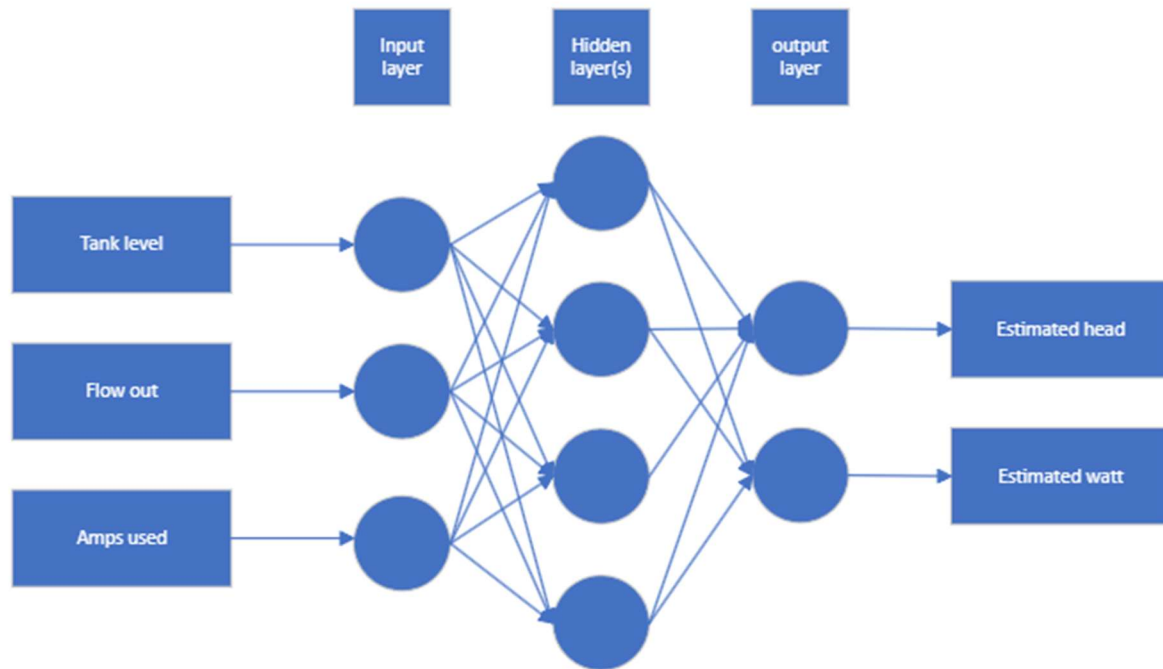


Figure 4-3 Neural network with alternative sensors as inputs with watt and head estimate as outputs.

4.8 Complementary tests for better insight

The tests will be conducted on the alternate sensor suite outlined in Section 4.4, utilizing a batch size of 16. The primary objectives of these tests are twofold: first, to explore potential avenues for enhancing the existing neural network model, and second, to gain deeper insights into the functioning of certain aspects of the network.

By analyzing the performance of the network under various conditions and configurations, we aim to identify areas where improvements can be made. Additionally, conducting these tests allows us to delve into the inner workings of the network, highlighting its behavior and providing valuable insights into its operations.

Through systematic experimentation and rigorous evaluation, we endeavor to refine the neural network model and uncover strategies for optimizing its performance. This iterative process of testing and analysis serves as a cornerstone for advancing our understanding of machine learning techniques and their applicability in real-world scenarios.

estimates

4.8.1 Tanh versus sigmoid activation function

Introducing variations in activation functions is a crucial step in refining neural network models. While sigmoid has been the primary activation function utilized thus far, it's imperative to explore other nonlinear functions to assess their efficacy. In this case, hyperbolic tangent (tanh) will be tested as an alternative to sigmoid.

By substituting the activation function with tanh and re-running the training process, we can compare its performance against the previously tested sigmoid function. This comparative analysis will provide valuable insights into the suitability of tanh for the given task and may reveal any advantages or drawbacks compared to sigmoid.

4.8.2 More data in the test dataset

Incorporating the 75% operating point into the test dataset instead of the training dataset presents an opportunity to evaluate the robustness and generalization capability of the neural network model. By introducing this change, we aim to assess how the inclusion of data points near the operating point affects the network's performance.

4.8.3 creating two new single output networks for head and watt

Transitioning from a neural network with two outputs (head estimate and watt estimate) to two separate networks, each with one output, can have several implications on the model's performance and computational efficiency.

In the previous setup, with two outputs in a single network, the model learns to simultaneously predict both the head estimate and watt estimate based on the input data. This approach allows for joint optimization of both tasks and potentially captures any dependencies or correlations between the two outputs.

However, splitting the model into two separate networks, each dedicated to predicting a single output, alters this dynamic. Each network focuses exclusively on one task, potentially leading to specialized models optimized for their respective outputs. This separation may offer advantages in terms of interpretability and modularity, as each network can be individually tuned and optimized for its specific task.

Figure 4-4 and Figure 4-5 illustrate the architectures of the new networks with only one output each. By comparing the performance of these networks with the previous dual-output network, we can evaluate the trade-offs and advantages of each approach. Factors such as training time, computational resources, and predictive accuracy should be considered when determining the most suitable architecture for the given task and dataset.

estimates

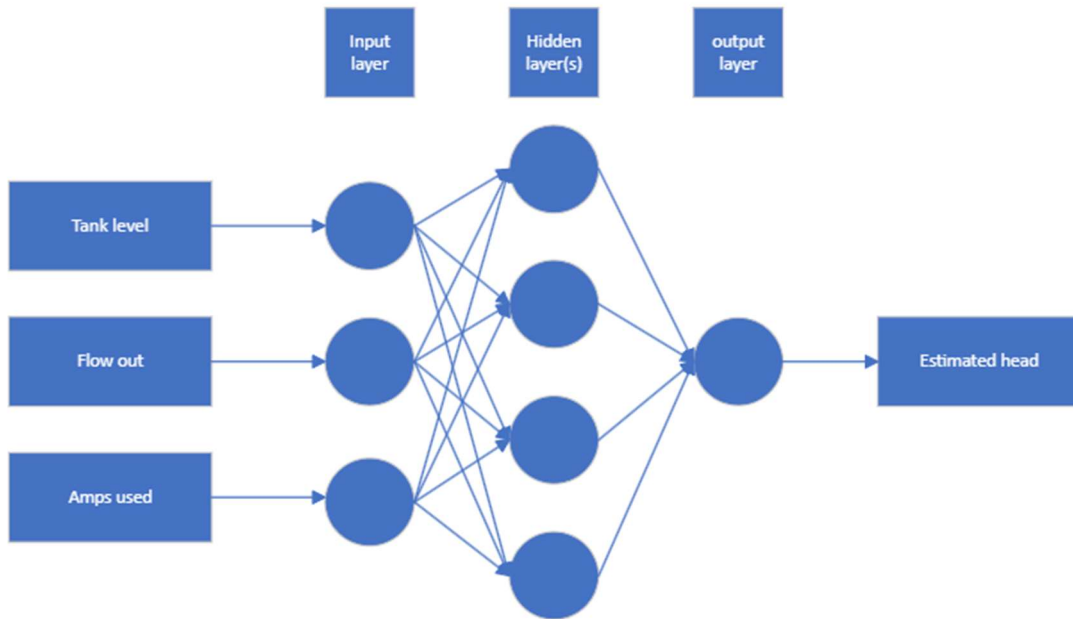


Figure 4-4 One output neural network with alternative sensors as inputs and head as output.

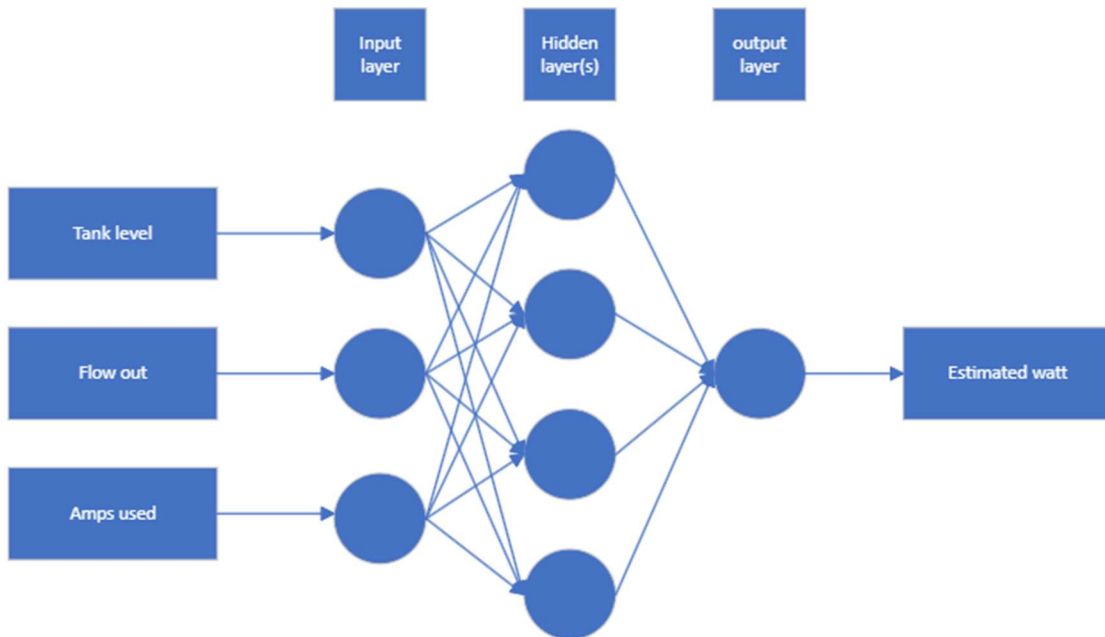


Figure 4-5 One output neural network with alternative sensors as inputs and watt as output.

estimates

4.8.4 Adding additional calculated features as inputs

Introducing additional features to the network can enhance its capability to capture complex relationships and improve predictive performance. In this case, the feature of $\frac{flow}{level}$ ratio is proposed as an additional input to the network, as depicted in Figure 4-6.

The flow/level ratio provides insight into the relationship between the flow rate of the pump and the level of the nearby tank. By incorporating this ratio as an input feature, the network gains access to additional information that may aid in better understanding and predicting the pump's behavior.

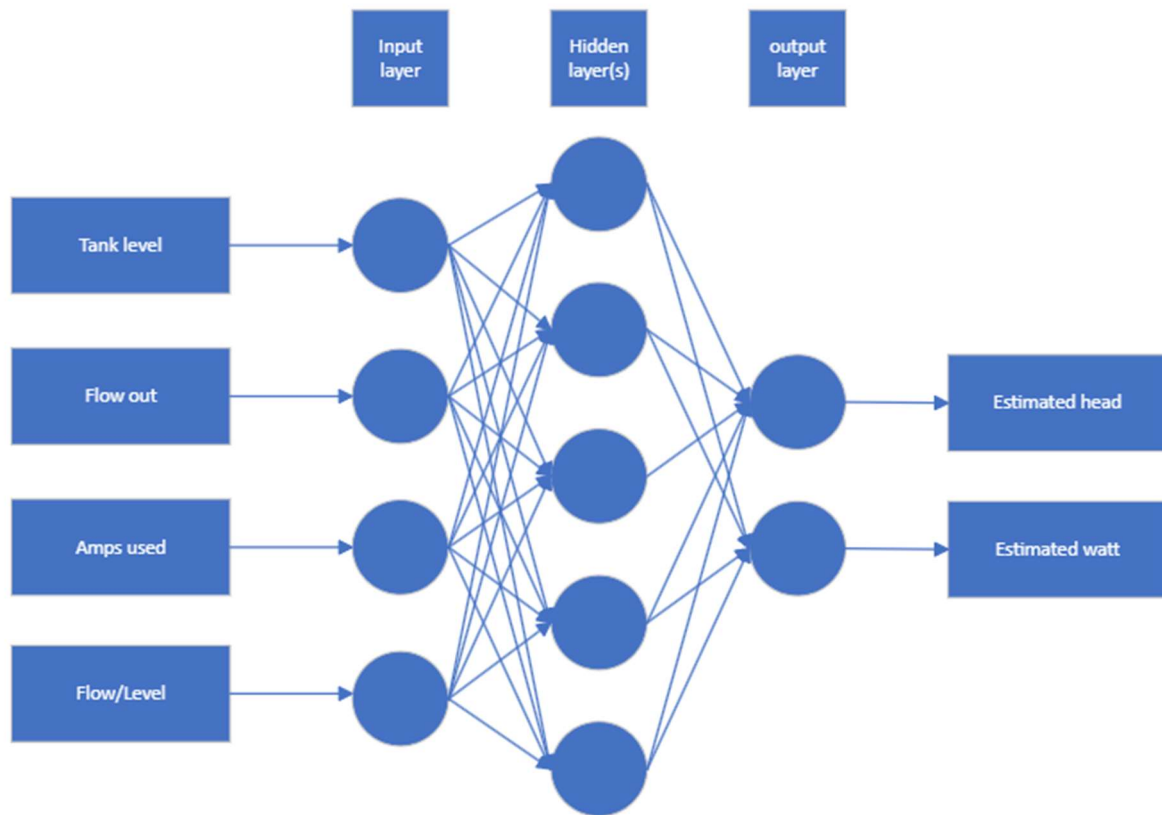


Figure 4-6 Neural network with alternative sensors and with one additional feature (flow/level) added. Outputs head and watt.

4.9 Reducing the network down to its basic weights and biases matrices

Equations (4-10) to (4-15) describe the operations of a 2-layer neural network, with (4-13) and (4-14) presenting their general forms. Despite the reference focusing on manual neuron estimation, these equations are applicable to equivalent networks. A trained neural network's functionality relies on matrix multiplication and addition, governed by its weights and biases. Once trained, the network can operate solely based on these matrices, enabling efficient deployment and prediction without access to the original data or network architecture. [20]

estimates

$$z_1 = W_1x + b_1 \quad (4-10)$$

$$a_1 = \text{sigmoid}(z_1) \quad (4-11)$$

$$z_2 = W_2a_1 + b_2 \quad (4-12)$$

$$a_n = f(z_n) \quad (4-13)$$

$$z_{n+1} = W_{n+1}a_n + b_{n+1} \quad (4-14)$$

$$\hat{y} = z_{n+1} \quad (4-15)$$

z_1, z_2 represent the output per network layer, W_1, W_2 are the network weights, b_1, b_2 are the biases, a_1 represents the layer z_1 after using the activation function and lastly \hat{y} is the output vector and x is the input vector.

With some straightforward code implementation, these parameters can be extracted from a trained model.

4.10 Transfer learning from p1001 to p1002

Transfer learning is the process of utilizing an already existing model and slightly modifying it to fit a similar but different process. In this case training it on p1001 as described earlier then adding on different extra layers to it and testing it. [19]

First test is a simple 2 extra output nodes, as seen in Figure 4-7

The second test is adding 2 more layers with 3 and 2 nodes. This one will also include a sigmoid activation function like in the original model, as seen in Figure 4-8

from knowledge gained from the other models, training it for 100 epochs is unnecessary so only 25 will be used here.

estimates

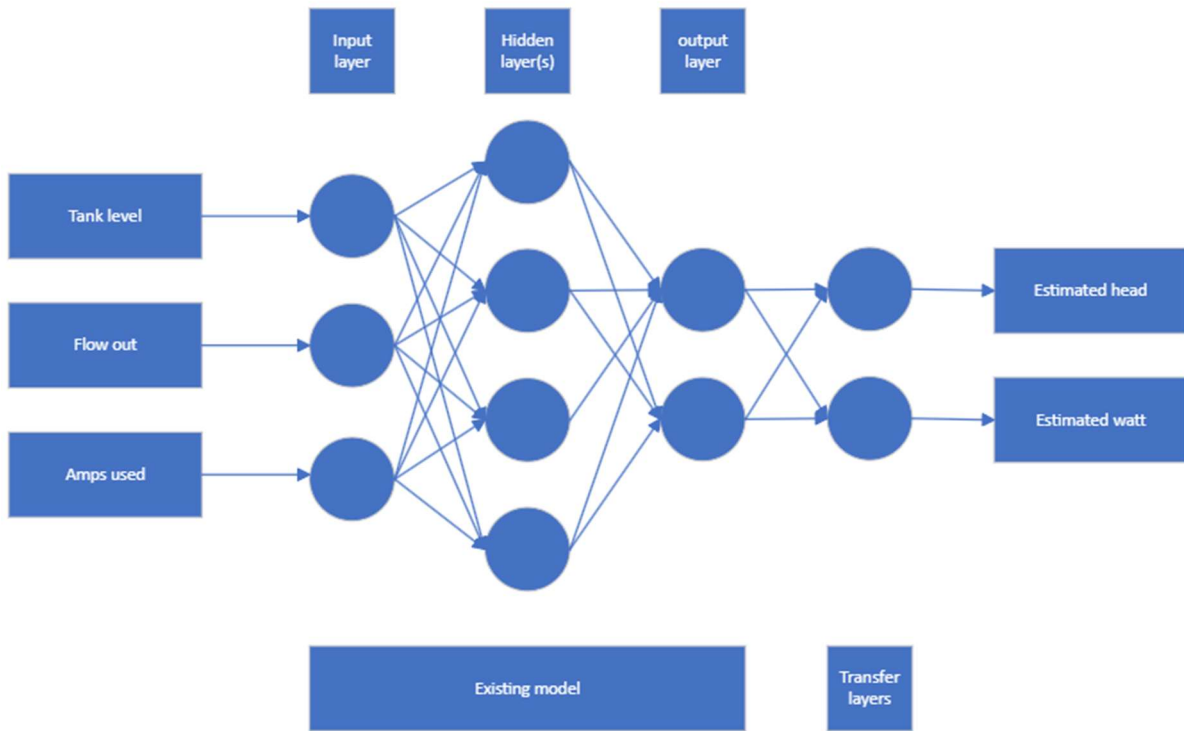


Figure 4-7 Existing alternate sensor neural network with an additional 2 nodes to be trained in transfer learning.

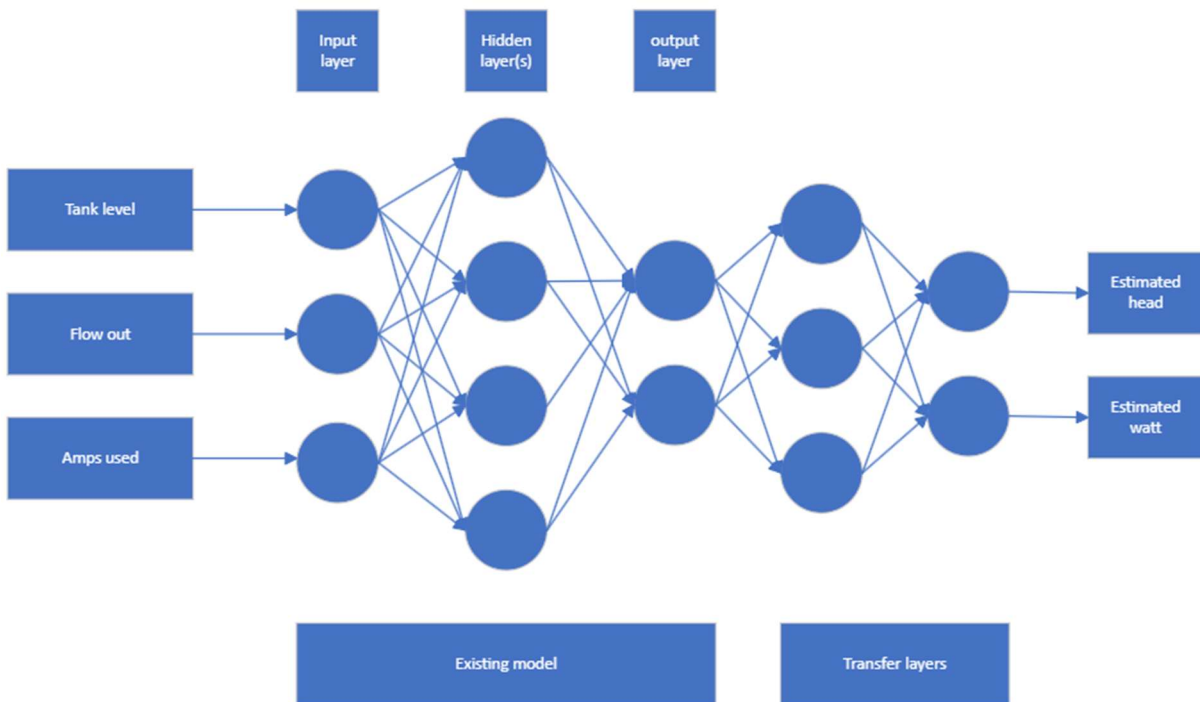


Figure 4-8 Existing alternate sensor neural network with an additional two layers to be trained in transfer learning.

5 Data analysis and statistics

This chapter will undertake an exploration and comprehensive exposition of findings derived from the methodologies outlined in Chapters 3 and 4.

In Appendix B, all code developed for this thesis, along with the accompanying CSV file containing the data, can be found in the GitHub repository. It's important to note that while the code itself won't be discussed in greater detail, the results generated from this code will be thoroughly examined and analyzed throughout the thesis. Readers interested in exploring the technical aspects further are encouraged to refer to the GitHub repository for access to the code and data.

5.1 Data gathered and pre-processed

Following the experimental design delineated in Section 3.2, a considerable volume of data was collected. With a 1 second resolution.

Concerning pre-processing, the raw data was initially stored in a series of CSV files. Notably, the control system intermittently opted to eliminate trends and configure the setup for data extraction, necessitating a mechanism to restore the columns to their correct order. This task was accomplished through a dedicated Python script leveraging the Pandas library. The resultant output yielded a unified CSV file consolidating all data points, subsequently serving as the primary dataset for all subsequent analyses.

5.2 Description of the Data

The data collected through the experimental design and the test rig will exhibit commonalities, which will be elucidated in this section and expounded upon in the ensuing subchapters. Subsequent sections will delineate several pertinent aspects, primarily focusing on the way the control system processes and archives data.

- Daca, data acquisition
- Pida, PID regulator
- Pv, Process value
- Op, operating point
- Speed, communicates with variable frequency drives for speed control.

All numerical values originate from physical input-output (IO) sources, with occasional exceptions where they are derived through calculations based on actual sensor readings or predetermined constants, serving the specific purpose of inclusivity within the dataset specifications.

Conducting a variable-by-variable examination entails scrutinizing the implications of the data and establishing connections between them. This analysis will abstain from delving into the statistical measures associated with each column and instead focus solely on elucidating the nature of the data. For comprehensive descriptions and column names, Table 5-1 serves as the repository.

5 Data analysis and statistics

Table 5-1 description of all variables and constants in the dataset.

Variable name	Description
PI1000.daca.pv	The pressure at the suction port P1001 demonstrates variability corresponding to the fluid level of the tank positioned proximally, denoted as B1001 (LI11001.daca.pv). This indicates a discernible correlation between the tank's fluid level and the suction pressure. Such behavior is typically anticipated in physical systems. However, it is important to note that the direct translation from level to pressure may not always be definitive. This uncertainty arises from factors such as piping layout and instrument placements. Even when the tank indicates 0% fluid level, residual fluid height may persist within the pipe. Thus, the installation of a pressure sensor at the suction port remains imperative for accurate monitoring and control.
P1001.speed.pv	The speed setting, expressed as a percentage of the variable frequency drive's range, necessitates validation against the configured parameters. In this instance, the range is specified as 0 to 50Hz.
PI1011.daca.pv	Pressure at outlet port pump p1001, represents a parameter subject to significant variation, primarily in response to changes in flow rate adhering to the pump curve. In this context, the observed behavior aligns with the anticipated performance characteristics.
FIC1003.pida.pv	The flow after the outlet from pump P1001 exhibits variability corresponding to PI1001, in accordance with the established pump curve relationship.
P1001.daca.pv	The current drawn by the pump, P1001, during operation, typically measured in amperes (A), is indicative of its electrical power consumption.
P1002.daca.pv	The current drawn by the pump, P1002, during operation, typically measured in amperes (A), is indicative of its electrical power consumption.
PI1013A.daca.pv	The pressure at the suction port of pump P1002 exhibits variation commensurate with the fluid level in tank B1002 and the control signal LIC1005, aligning with anticipated behavior.

5 Data analysis and statistics

PI1013B.daca.pv	The pressure measurement at the outlet of pump P1002 has been observed to present challenges, predominantly attributable to external factors beyond the control system and the machinery it interfaces with. Notably, this sensor exhibited minimal variation despite significant alterations within the system, suggesting potential external influences on its performance.
FI1015.daca.pv	The flow discharged from pump P1002 experienced cessation when the control valve closed beyond a threshold of 45%, contributing to the challenges associated with the operation of pump P1002.
P1002.speed_sp.pv	Essentially identical to p1001.speed.pv, the speed of pump P1002 followed a similar pattern; however, the variable frequency drive was configured differently. Consequently, this pump operated consistently at 60%, aligning with the surrounding dataset observations.
TI1011A.daca.pv	The ambient temperature at the location of the test rig refers to the prevailing temperature conditions in the immediate environment surrounding the experimental setup.
Lic1005.pida.op	The operating point of the level controller denotes the specific value representing the opening of the valve subsequent to pump P1002 in this context.
Fic1003.pida.op	The operating point of the flow controller denotes the specific value representing the opening of the valve subsequent to pump P1001 in this context.
LI1001.daca.pv	The level of tank B1001
Li1005.pida.pv	The level of tank B1002
Volt, p1001.Q_min, p1001_Q_max, p1001_Head, p1002.Q_min, p1002.Q_max, p1002.head	These columns hold the system voltage, and the pumps specification as min flow, max flow, and rated head.

5 Data analysis and statistics

H1	Calculated Pump head for p1001 with equation (2-2)
H2	Calculated pump head for p1002 with equation (2-2)
Watt1	Calculated watt used by p1001 with equation (2-5)
Watt2	Calculated watt used by p1002 with equation (2-5)

5.3 Data statistics

Table 5-2 exhibits the variable names arranged as rows, accompanied by their respective mean values and standard deviations. It encompasses the entirety of the dataset, inclusive of components not utilized in the machine learning process or curve fitting analyses.

Table 5-2 statistics for each of the variables in the full dataset showing mean value, standard deviation, and unit.

Measurement point	Mean	Standard deviation	unit
PI1000.daca.pv	48.40	7.35	mBar
P1001.Speed.pv	100	0	%
PI1011.DACA.PV	952.56	42.79	mBar
FIC1003.PIDA.PV	11.16	5.33	L/min
p1001.daca.pv	1.32	0.04	A
p1002.daca.pv	1.32	0.03	A
PI1013A.DACA.PV	79.08	30.4	mBar
PI1013B.DACA.PV	697.46	33.02	mBar
FI1015.DACA.PV	12.33	7.34	L/min
P1002.Speed_SP.pv	60	0	%
TI1011A.DACA.PV	21.48	0.01	°C
lic1005.pida.op	80.54	18.39	%
fic1003.pida.op	60.36	25.01	%
LI1001.DACA.PV	26.40	6.29	%
lic1005.PIDA.PV	54.47	16.97	%
Volt	230	0	V
p1001.Q_min	10	0	L/min
p1001.Q_max	50	0	L/min
p1001.head	14	0	m

5 Data analysis and statistics

p1002.Q_min	10	0	L/min
p1002.Q_max	50	0	L/min
p1002.head	14	0	m
H1	9.21	0.42	m
H2	6.30	0.1	m
Watt1	364.16	12.09	W
Watt2	364.08	9.99	W

Column H1 is derived from PI1000, PI1011, and FIC1003.pv, all of which exhibit a noteworthy standard deviation. This suggests that the dataset associated with Column H1 is likely to yield valuable insights and produce favorable outcomes in both machine learning and curve fitting endeavors.

Conversely, Column H2 is computed from PI1013A, PI1013B, and FI1015, displaying a low standard deviation. Consequently, it is anticipated that this dataset will not perform optimally in machine learning or curve fitting tasks. This limitation stems from the polynomial nature of a pump curve, where a low variance may result in insufficient learning or the propensity to converge on a singular output value.

Columns characterized by a standard deviation of 0 denote constant values utilized for alternative purposes within the code.

The higher degree of standard deviation observed in PI1013A, compared to PI1000, can be attributed to the differential heights of their corresponding tanks, B1002 and B1001, respectively. Tank B1002 is approximately three times taller than tank B1001, leading to a broader range of potential values. This disparity in tank heights influences the variability in pressure readings, which is expected to manifest more prominently in subsequent machine learning algorithms.

5.4 Student t-test

Setting up a t-test for the training data and test data has yielded further insights. Employing a significance level of 0.05 to assess the similarity between columns, it was observed that within the P1001 dataset, two columns exhibited p-values of 0.07 and 0.24, flow FIC1003 and head H1, indicating a level of similarity deemed slightly concerning Table 5-3. Conversely, in the dataset for P1002, no columns displayed similarity to such an extent Table 5-4

Table 5-3 P1001 t test results showing t statistic, p value for each of the variables level flow, amps, head, and watt.

	LI1001	FIC1003	P1001	H1	Watt1
t-statistic	12,70	1,81	6,21	1,17	6,21
p-value	8,52e-17	7e-2	5,17e-10	2,41e-1	5,16e-10

Table 5-4 P1002 t test results showing t statistic, p value for each of the variables level flow, amps, head, and watt.

	LiC1005	FI1015	P1002	H2	Watt2
t-statistic	-6,12	38,02	17,95	-33,67	17,95
p-value	9,53e-10	3e-304	1,88e-74	4,78e-241	1,88e-71

Note, in this data Level, flow and amps are inputs to the neural network while head and watt are the outputs.

5.5 Data correlation

Understanding the correlations between different aspects of the dataset is crucial for gauging its success. The figures below illustrate the correlations between columns in the full dataset, as well as the training data for P1001 and P1002, respectively. Positive correlations are depicted in blue, negative correlations in red, with darker shades indicating stronger correlations. Lack of correlation is represented by white areas.

Figure 5-1 displays the full dataset, excluding constant values. Given the presence of 18 variables in this dataset, the figure may present challenges in readability. It's important to note that both pumps are included in this figure, although they wouldn't be simultaneously utilized in machine learning endeavors.

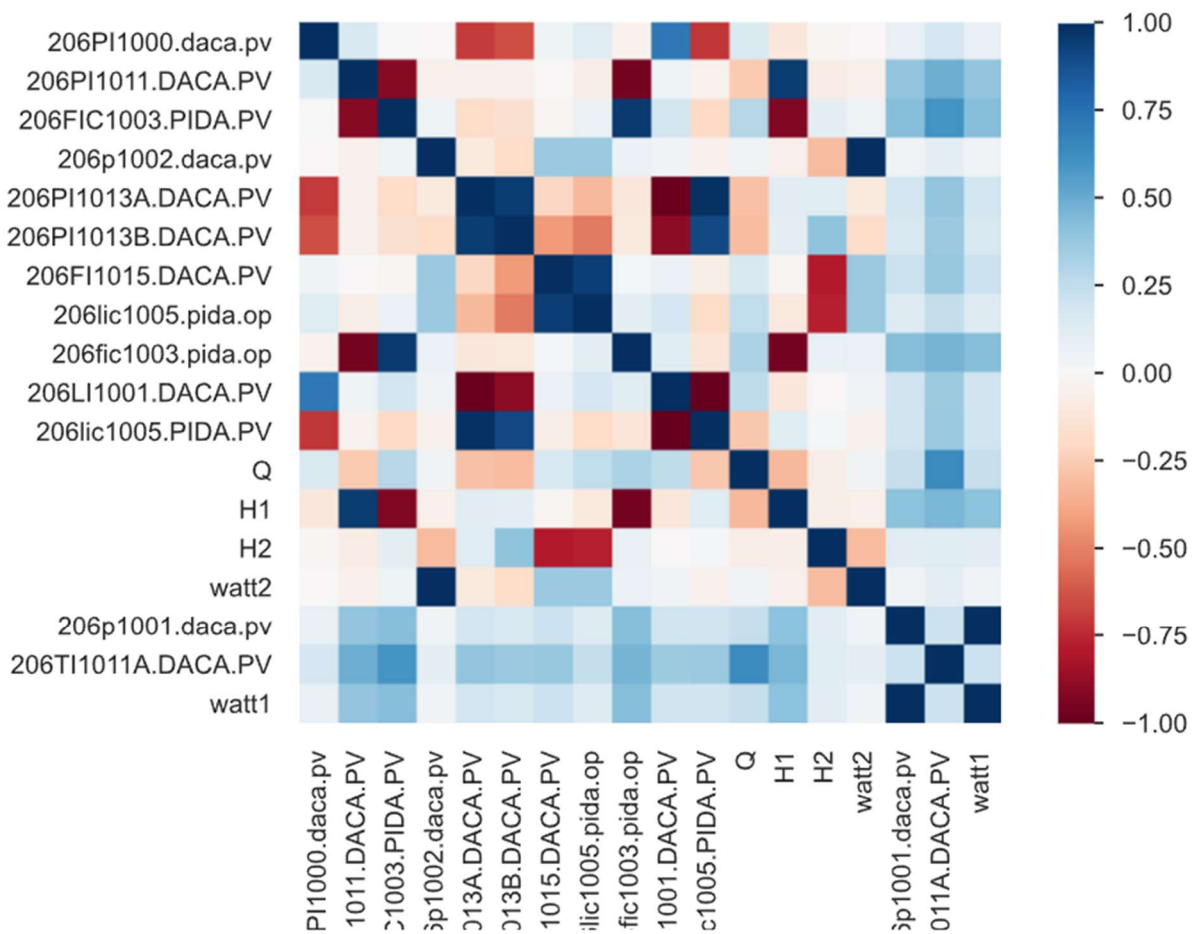


Figure 5-1 Correlation matrix for the full dataset.

Figure 5-2 illustrates the correlation within the training data for P1001, with the corresponding correlation values provided in Table 5-5. The columns H1 and Watt1 serve as the y vectors for machine learning.

One point of interest is the level indicator, LI1001, which exhibits weak correlations with every other column. This suggests that LI1001 contains information that can be effectively utilized in conjunction with other columns to derive conclusions within an algorithm.

Additionally, the flow indicator, FIC1003, displays a strong negative correlation with the head (H1). This correlation aligns with expectations when considering the pump's H-Q curve,

5 Data analysis and statistics

indicating that flow contains significant information regarding the pump's head. However, flow does not exhibit as strong a correlation with the power used by the pump/motor (P1001 and Watt1).

Moreover, the power drawn by the pump demonstrates a perfect 1-to-1 correlation with Watt1. This correlation is logical considering how Watt1 is calculated, with the only measured input value being the motor's amperage. Thus, a high correlation in this context is expected and makes sense.

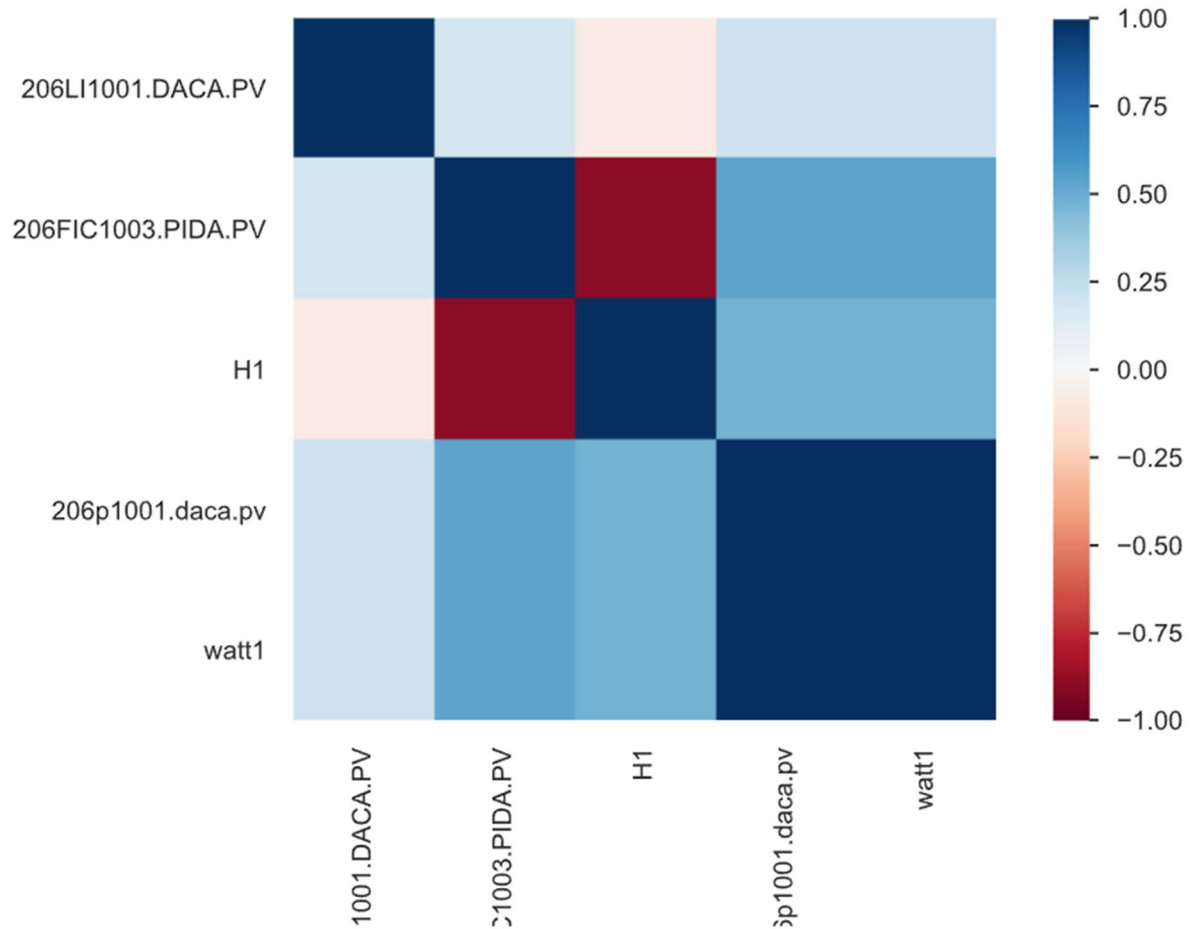


Figure 5-2 Correlation matrix for the training data for p1001. Alternative sensors suit with level, flow, calculated head, pumps amp drawn, and calculated watt shown as tags.

5 Data analysis and statistics

Table 5-5 Correlation for the training data for p1001

	Li1001	Fic1003	H1	P1001	Watt1
Li1001	1	0.184	-0.08	0.21	0.21
Fic1003	0.184	1	-0.894	0.526	0.526
Hi	-0.08	-0.894	1	0.473	0.473
P1001	0.21	0.526	0.473	1	1
Watt	0.21	0.526	0.473	1	1

Figure 5-3 illustrates the correlation within the training data for P1002, with corresponding correlation values provided in Table 5-6. The columns H2 and Watt2 serve as the y vectors for machine learning.

Upon comparison with the P1001 data, notable differences emerge. Firstly, the correlation between the level indicator, LIC1005, and other columns is closer to 0, indicating that the level of tank B1002 exerts less influence on other parts of this dataset. This may potentially impede the effectiveness of machine learning models.

Furthermore, the flow (FI1015) and head (H2) exhibit a weaker negative correlation compared to the P1001 data (-0.894 in P1001 and -0.368 in P1002). This weaker correlation suggests potential challenges in achieving high accuracy in machine learning models for P1002 data.

Similarly, the amps (P1002) and power (Watt2) demonstrate similar patterns to those observed in P1001.

5 Data analysis and statistics

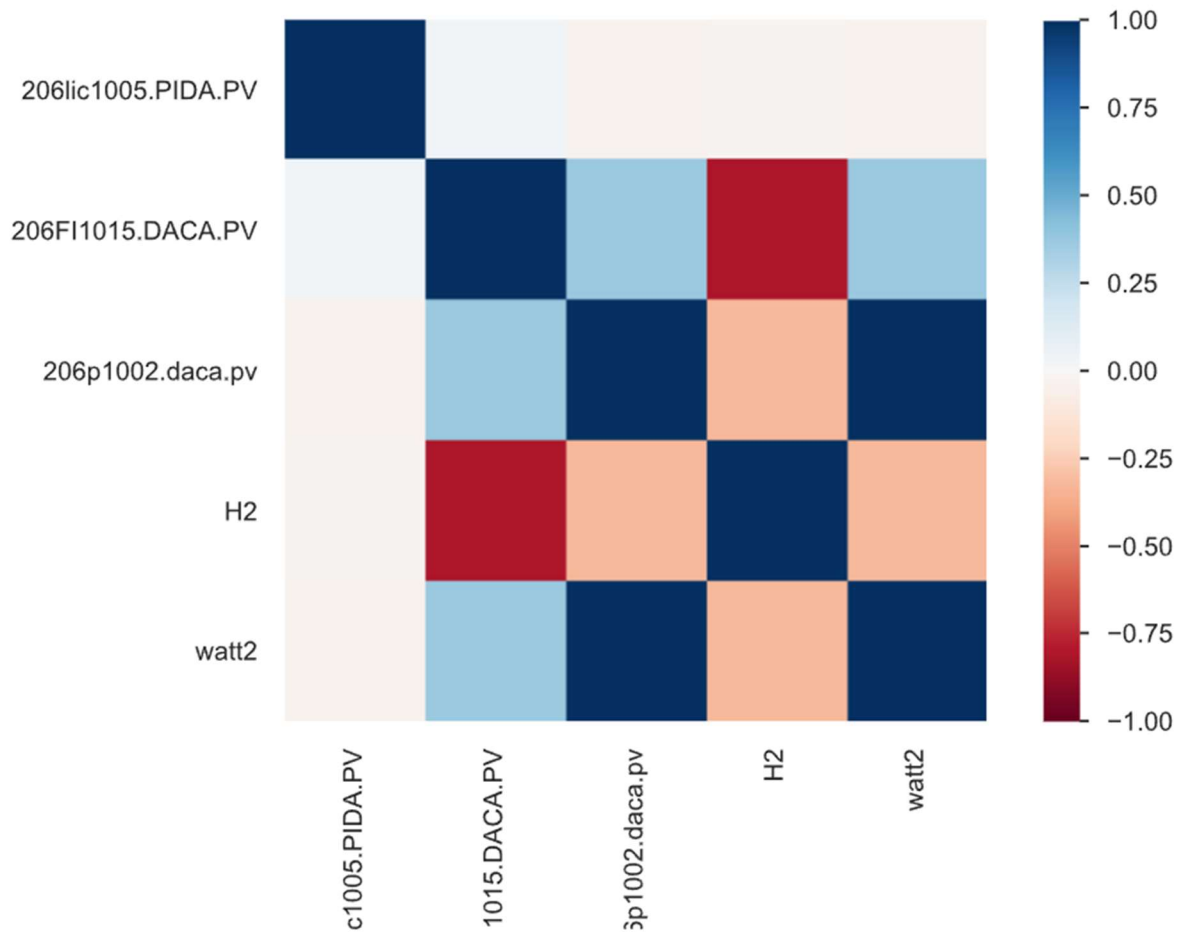


Figure 5-3 Correlation for the training data for p1002. Alternative sensors suit with level, flow, pumps amp drawn, calculated head, and calculated watt shown as tags.

Table 5-6 Correlation for the training data for p1002

	Lic1005	Fi1015	P1002	H2	Watt2
Lic1005	1	0.026	-0.036	0.025	-0.036
Fi1015	0.026	1	0.368	-0.81	0.368
P1002	-0.036	0.368	1	-0.327	1
H2	0.025	-0.81	-0.327	1	-0.327
Watt2	-0.036	0.368	1	-0.327	1

5.6 Analytical H-Q curve

The mathematical calculations employed to generate this curve were detailed in Chapter 4.1. From these calculations, the H-Q curve emerges, delineating the operational preferences of the pump itself.

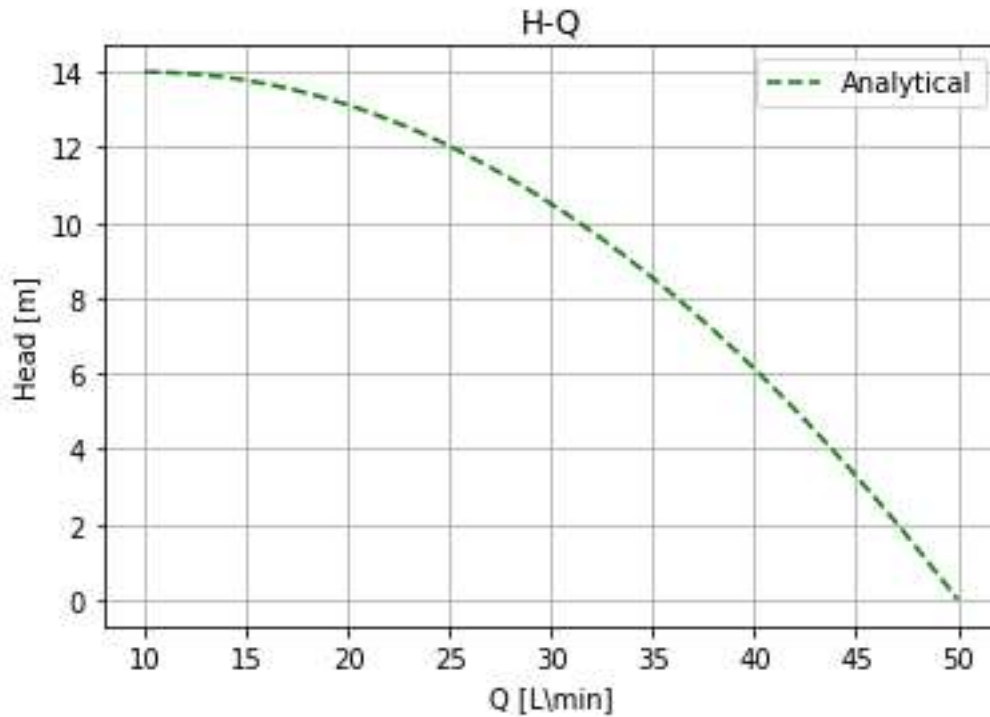


Figure 5-4 Estimated HQ Curve made from hand calculating b_p and known H_{max} on a 10 to 50 L/min flow rate x axis.

5.7 Data driven H-Q curve from data gathered

Given that this system comprises similar pumps operating in two distinct configurations, it's imperative to interpret the data for each pump within the context of its respective surroundings.

5.7.1 Data driven H-Q curve for p1001

Initiating the analysis with P1001 following tank B1001, as depicted in Figure 5-5, we observe that the maximum pump head is expected to be 14m. However, even when operating the pump against a closed valve, it only reached 10m. The valve in question, positioned approximately 4 meters above ground level, could influence the reading at 0 L/min due to losses attributed to hydrostatic head. Unfortunately, this influence was not considered in the original data preprocessing.

Despite this, the curve effectively fits the data in a manner consistent with the anticipated curve. System constraints dictate that the pump cannot exceed a flow rate of 18 l/min.

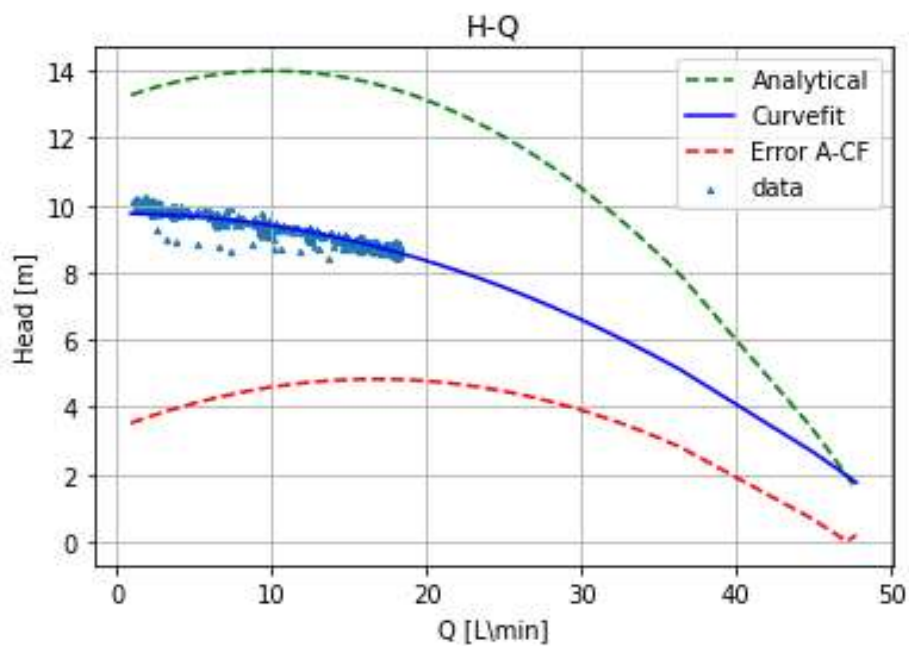


Figure 5-5 P1001 analytical in green, curve fitted curve in solid blue and error curve in red with data as blue triangles.

5.7.2 Data driven H-Q curve for P1002

Moving on to P1002, which experiences a loss of 6m due to hydrostatic pressure and consistently fails to achieve its maximum rated head, regardless of experimental conditions. In this scenario, the curve fitting process aiming to match a particular pump curve to its data did not yield the expected results. This discrepancy underscores the significance of system variations surrounding the pump, illustrating how these factors significantly influence its performance. Figure 5-6

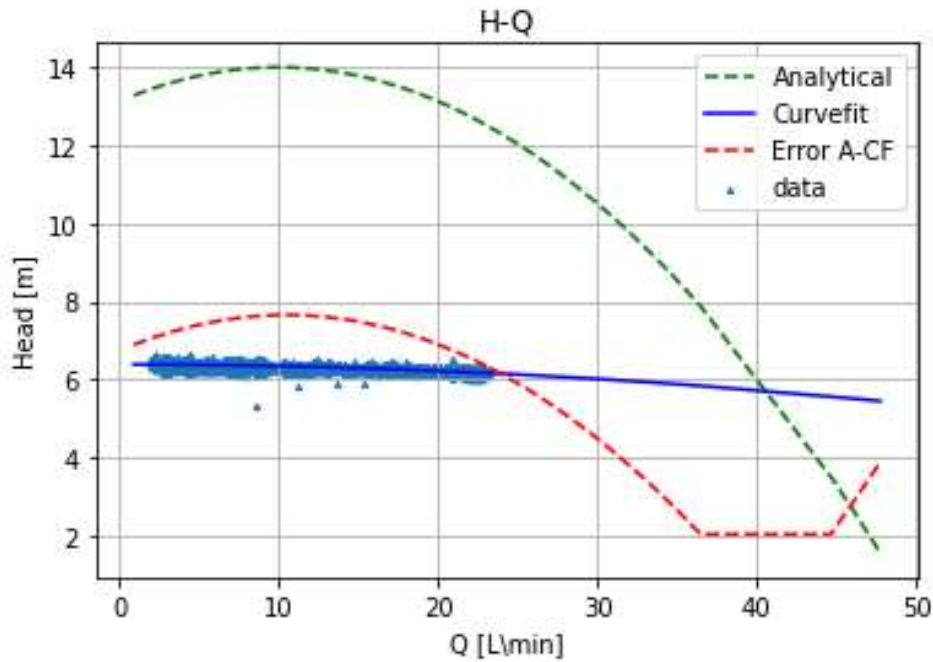


Figure 5-6 P1002 analytical in green, curve fitted curve in solid blue and error curve in rad with data as blue triangles.

5.7.3 Parameter values for both pumps

The values obtained for \hat{H}_{max} and \hat{b}_p are documented in Table 5-7. These represent the optimal values derived for the curve fitting process. The disparity observed in the \hat{b}_p term elucidates why the curve for p1002 is relatively flat in comparison to that of p1001.

Table 5-7 \hat{H}_{max} and \hat{b}_p estimated values for p1001 and p1002 from the curve fitting process. (table on two pages!!)

	\hat{H}_{max}	\hat{b}_p
P1001	9.75	0.0035
P1002	6.39	0.0004

6 Machine learning results

This chapter presents the findings from chapter 4. Commencing with application of the random forest regressor as a baseline for comparison, subsequent analysis delves in various neural network architecture and additional tests. Due to substantial distinctions in results between the two pumps. Their respective outcomes will be described in separate subchapters.

6.1 Random forest regressor head estimate

Upon configuring a random forest regressor (RFR) comprising of 100 trees and training them on the alternative sensor suit associated with p1001 and p1002, the outcomes are shown in Table 6-1.

The hyperparameters employed in the model were primarily set to default values provided by the scikit-learn Python library. Two parameters, however, were explicitly adjusted. The `n_estimators` parameter was set to 100, specifying the utilization of 100 trees within the regressor ensemble. Additionally, the `random_state` parameter was specifically configured to 42 to ensure deterministic behavior, facilitating consistent results across multiple runs. It's noteworthy that parameters such as `max_depth`, `min_samples_split`, and `min_samples_leaf` retained their default values, at None, 2, 1 respectively, as their adjustment was deemed unnecessary for the current modeling context.

The result of P1001 indicates its operability under favorable data conditions. Manifesting in a correlation confidence of 94.73%, the results align well within the anticipated range.

However, in the case of P1002 presents a contrasting scenario, its performance is so far unsatisfactory both in terms of curve fitting seen in previous chapter and within the framework of the random forest regressor. Due to the low r^2 scores of 55.20%. This can however be due to the high degree of variance withing the testing data, and not due to a faulty algorithm.

Table 6-1 MSE and r^2 scores from the random forest regressor for P1001 and P1002

Pump	Mean Squared Error	r^2 scores
P1001	0.05	94,73%
P1002	0,44	55.20%

6.2 Neural network head estimate

These findings are utilizing the normal sensors suit. This includes pressure at suction port, pressure at discharge port, flow out of the pump and amperes drawn by the pump.

Upon configuring the NN (neural network) depicted in Figure 4-1 with input data identical to that utilized for calculating the y vector, the ensuing results are as follows.

6.2.1 NN for P1001

For p1001 the findings were excellent, training for only 10 epochs with batch size 16 gave an R^2 score of 99.21%, more results seen in Table 6-2 and its related plot showing the training and testing loss and r^2 scores over 100 epochs in Figure 6-1.

Table 6-2 r^2 scores for p1001 at batch size 16, test score. Over select epochs up to 100

Batch size/ epochs	16
10	99,21%
25	99,55%
50	99,68%
100	99,82%

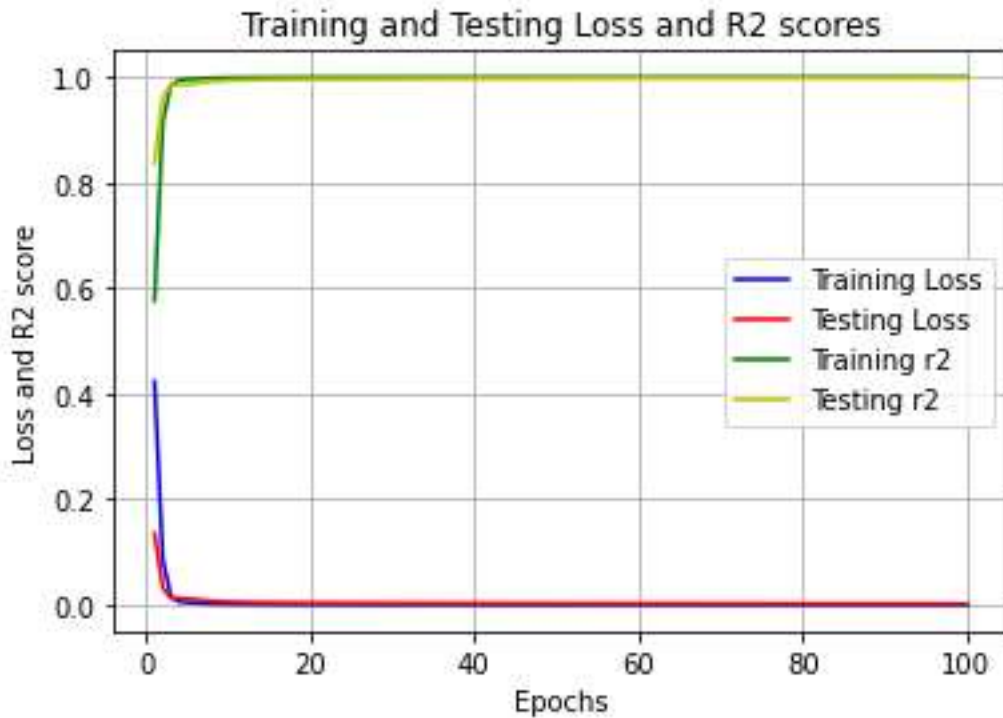


Figure 6-1 p1001 training and test loss graph with r^2 scores for p1001 normal sensor dataset.

Thus far, the examination has solely focused on the training phase of the neural network. However, it is imperative to extend the analysis to encompass the performance of the network when presented with validation data, in this case the validation data is the same as the test data. This entails plotting the networks' predictions against the validation data itself. This procedure will be repeated for all subsequent tests.

Figure 6-2 depicts the juxtaposition of the validation data represented in blue against to the predictions in orange, against the formerly calculated H-Q curve from the pump. As the network is so accurate there is barely any blue dots, hence why Figure 6-3 is zoomed in on the data itself to better clarify predictions versus validation data, resulting in an almost perfect fit.

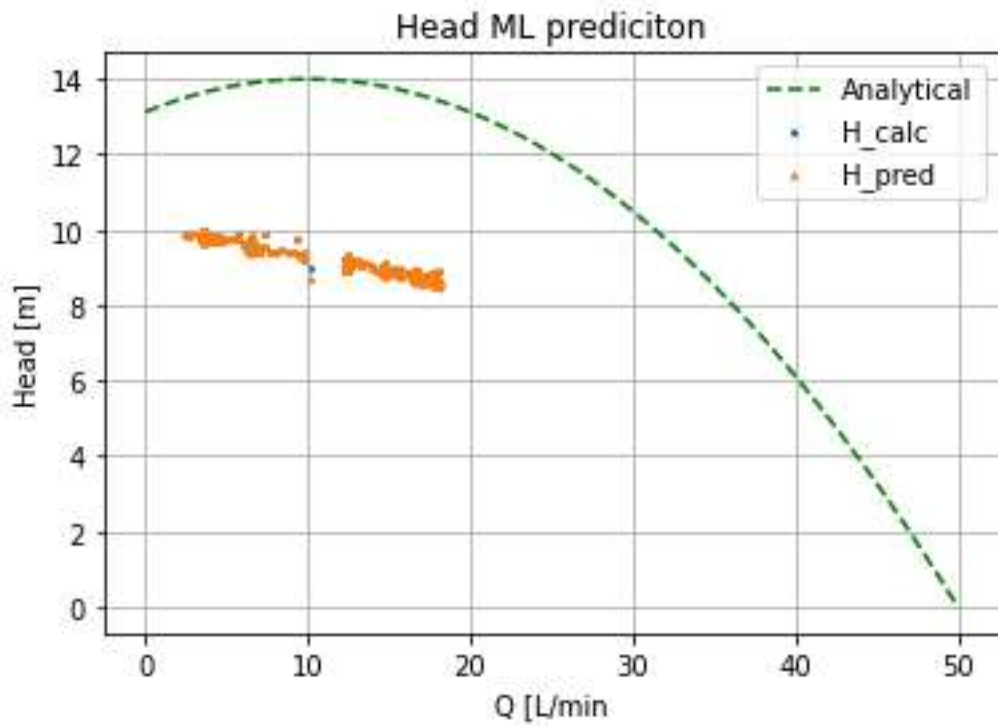


Figure 6-2 Original test data and predictions compared to analytical curve for P1001 normal sensors.

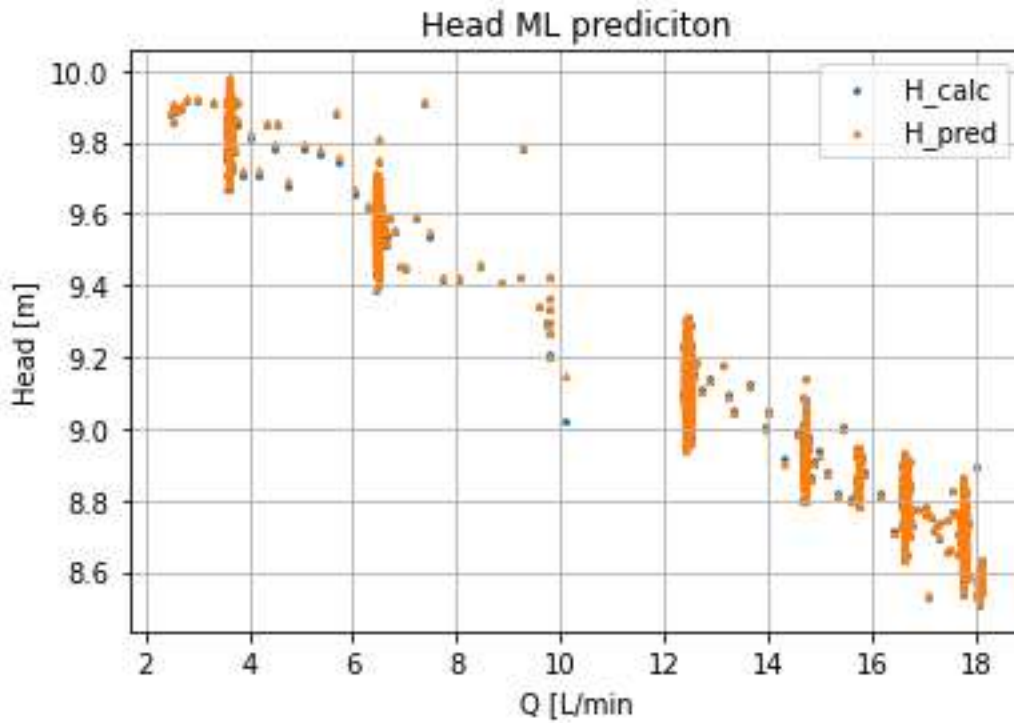


Figure 6-3 Original test data and predictions closer look for P1001 normal sensors.

6.2.2 NN for P1002

Much of the same can be seen for P1002, excellent r^2 scores after 10 Epochs, Table 6-3 and Figure 6-4. This high score contrasts the alternative sensor pack used in the random forest regressor from 6.1 and this is due to the sensors being used as input, being the same as the one used to calculate the output in the first place.

Table 6-3 r^2 scores for p1002 at batch size 16 Over select epochs up to 100.

Batch size/ epochs	16
10	99,18%
25	99,73%
50	99,86%
100	99,93%

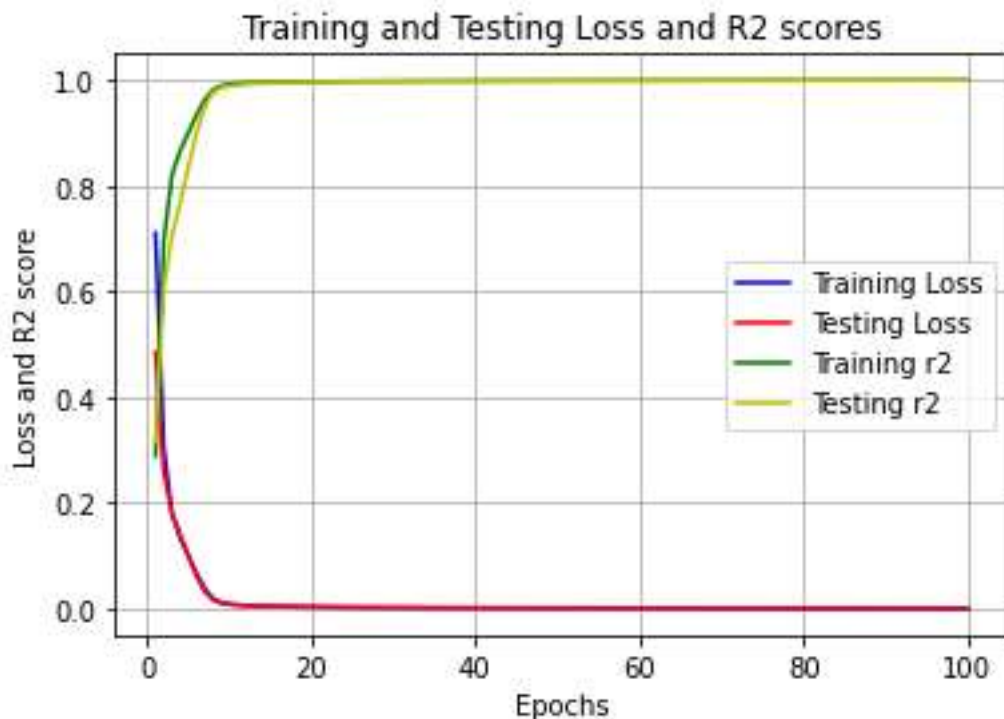


Figure 6-4 p1002 training and test loss graph with r^2 scores for p1002 normal sensor dataset.

Training of the neural network must be followed by other tests.

Figure 6-5 depicts the juxtaposition of validation data in blue against the prediction in orange and the previously calculated H-Q curve. The almost perfect overlap means Figure 6-6 is needed to highlight where the blue dots are located. In contrast to the data seen from Figure

6 Machine learning results

6-3, this data has a much higher degree of noise. This noise can be one of the contributing factors for the RFRs (random forest regressor) low accuracy and is caused by the high variance observed in PI1013B after tank B1002. The high variance will be a reoccurring phenomenon in subsequent chapters relating to P1002.

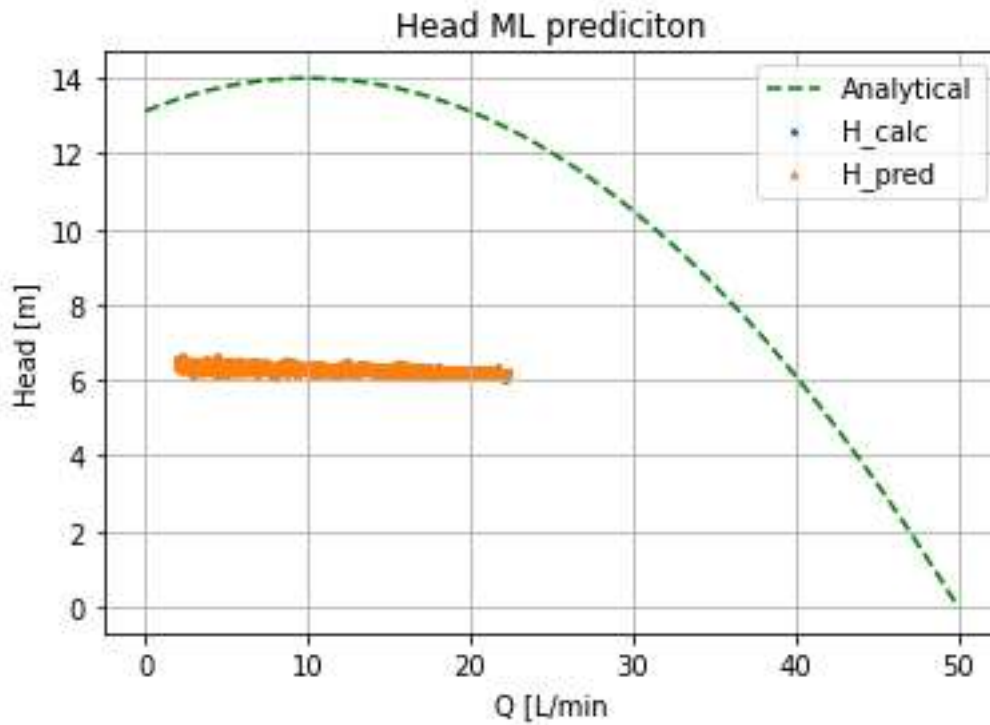


Figure 6-5 Original test data and predictions compared to analytical curve for P1002 normal sensors.

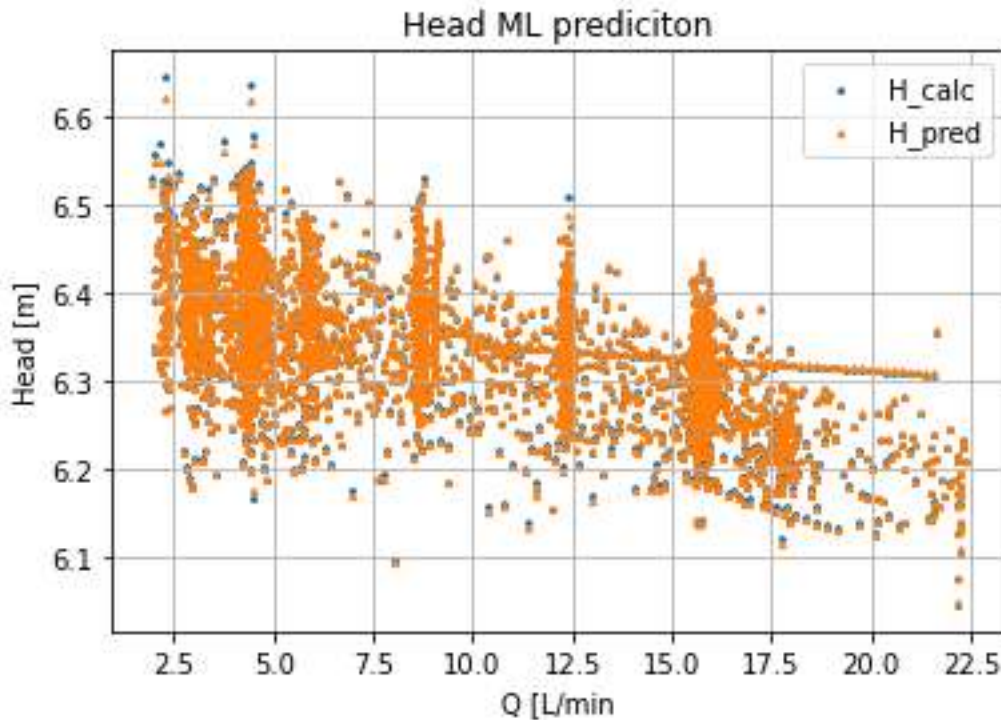


Figure 6-6 Original test data and predictions closer look for P1002 normal sensors.

6.3 Neural network head estimates using alternative sensors and batch size

These findings are utilizing the alternative sensor suit. This includes level of the tank before the pump, flow out of the pump and amperes used by the pump.

Upon configuring the neural network depicted in Figure 4-3 with input data different to that utilized for calculating the y vector, the ensuing results are as follows. Performing the tests varying the batch size hyperparameter from 16 to 64

6.3.1 NN alternative sensors P1001

For p1001 the results were once again excellent with the best result being 100 epochs at batch size 16, giving an r^2 scores of 98.35%, as seen from Table 6-4.

Figure 6-7 shows r^2 scores and training loss over the epochs, no overfitting or diverging test losses to be seen.

All this shows promising results to be utilized in transfer learning later.

6 Machine learning results

Table 6-4 r^2 scores of different training method around p1001, test score. Over select epochs up to 100

Batch size/ epochs	16	32	64
10	97.08%	97.0%	94.9%
25	97.78%	97.5%	96.9%
50	97.94%	97.9%	97.4%
100	98.35%	98.2%	97.8%

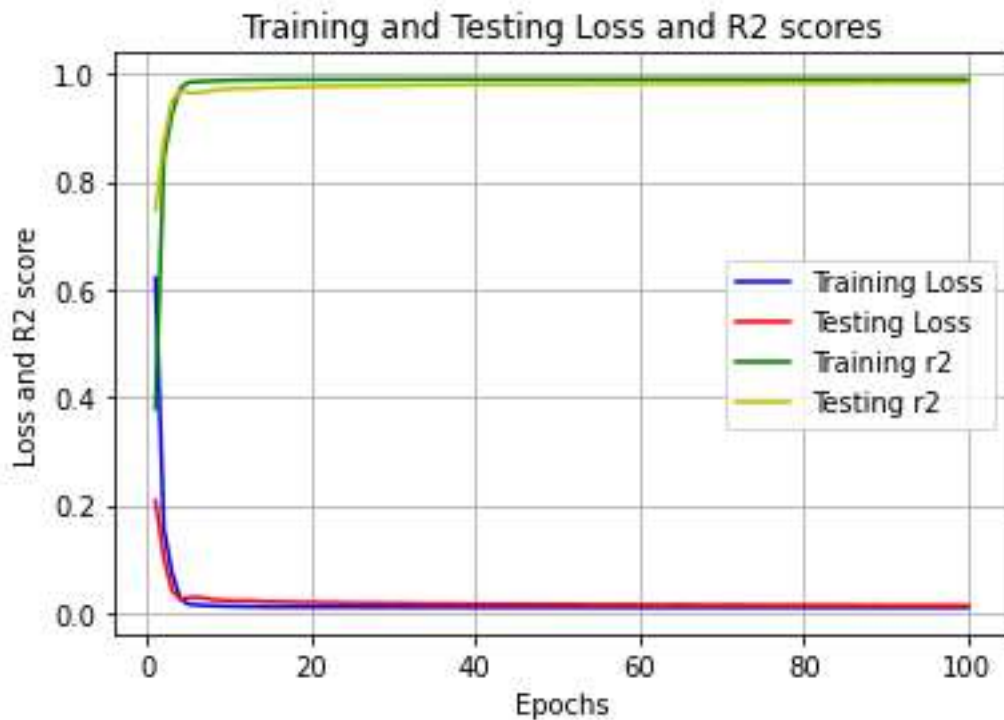


Figure 6-7 p1001 alternative sensors batch size 16 training and test loss graph with r^2 scores.

Figure 6-8 shows the prediction results in orange against the test data in blue against the analytical H-Q curve, bluer can be seen in this example than from the previous subchapter.

Figure 6-9 shows the same but more zoomed in, in this case it can really seem like the network has only picked up on the H-Q curve itself and not overfitted to find every datapoint possible. If the curve fitted curve was overlayed onto this, it would match well with the ML model prediction.

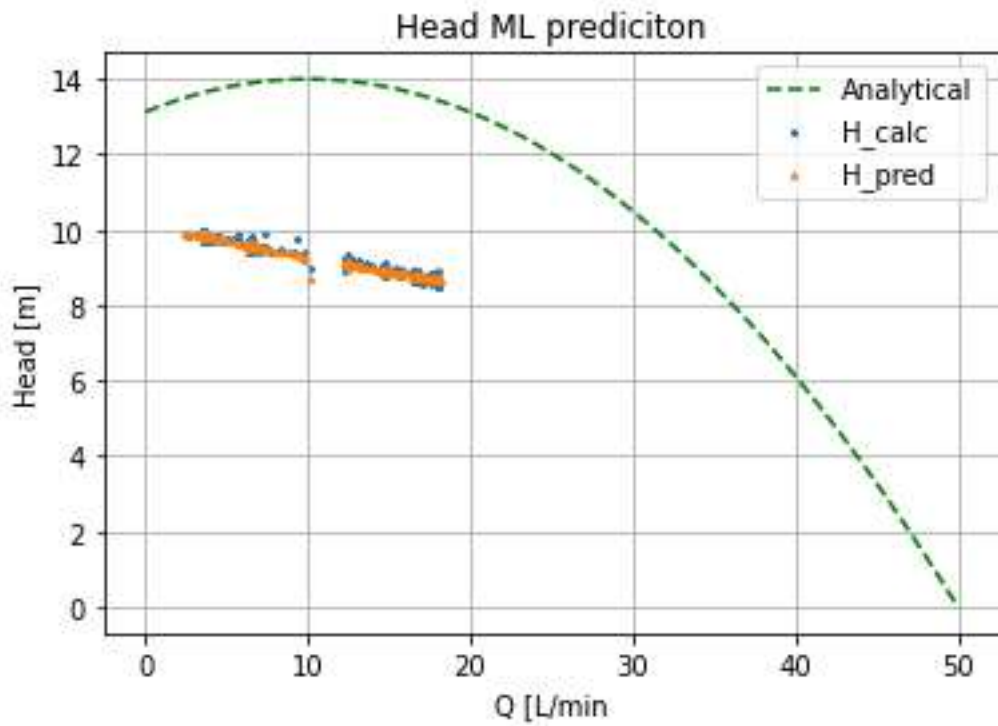


Figure 6-8 Original test data and predictions compared to analytical curve for P1001 alternative sensors.

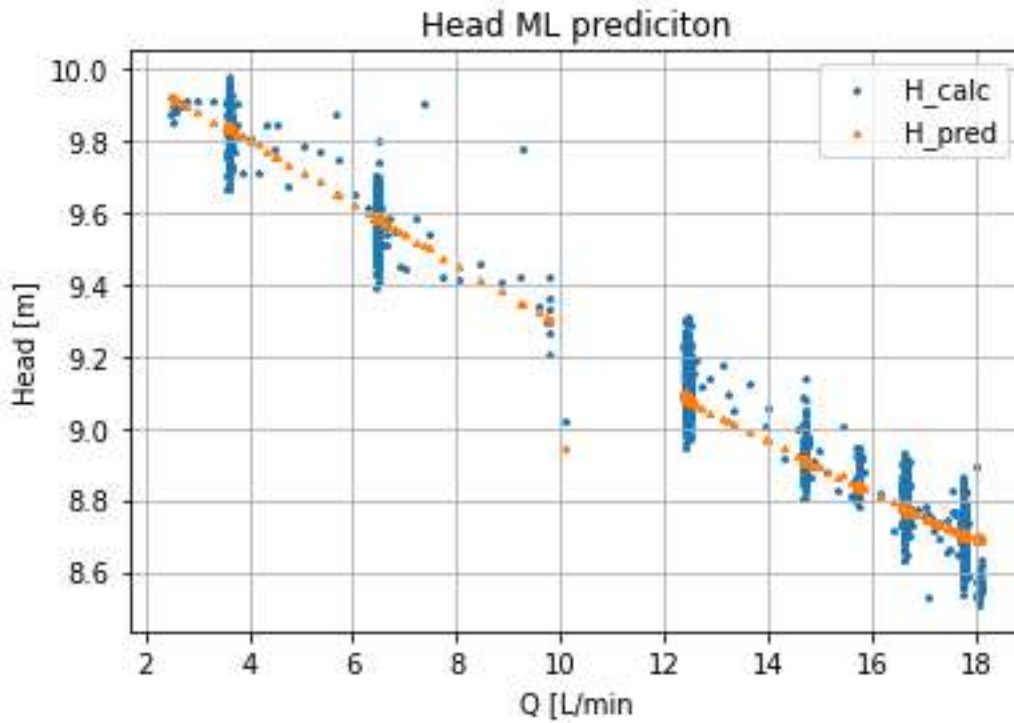


Figure 6-9 Original test data and predictions closer look for P1001 alternative sensors.

6.3.2 NN alternate sensors P1002

Performing the same training with the data from P1002 yields a different outcome compared to training with the normal sensor pack, aligning more closely with the RFR results. With its best training R^2 score reaches only 68%, Table 6-5 and the training converging early to that score. It appears that this setup for pumps does not yield favorable results. Figure 6-10 shows the training and test losses as well as r^2 scores.

Table 6-5 r^2 scores with different training methods around p1002, test score. Over select epochs up to 100

Batch size/ epochs	16	32	64
10	66,6%	66,1%	54,93%
25	67,9%	67,5%	66,1%
50	68,0%	67,9%	67,4%
100	68,0%	68,2%	67,9%

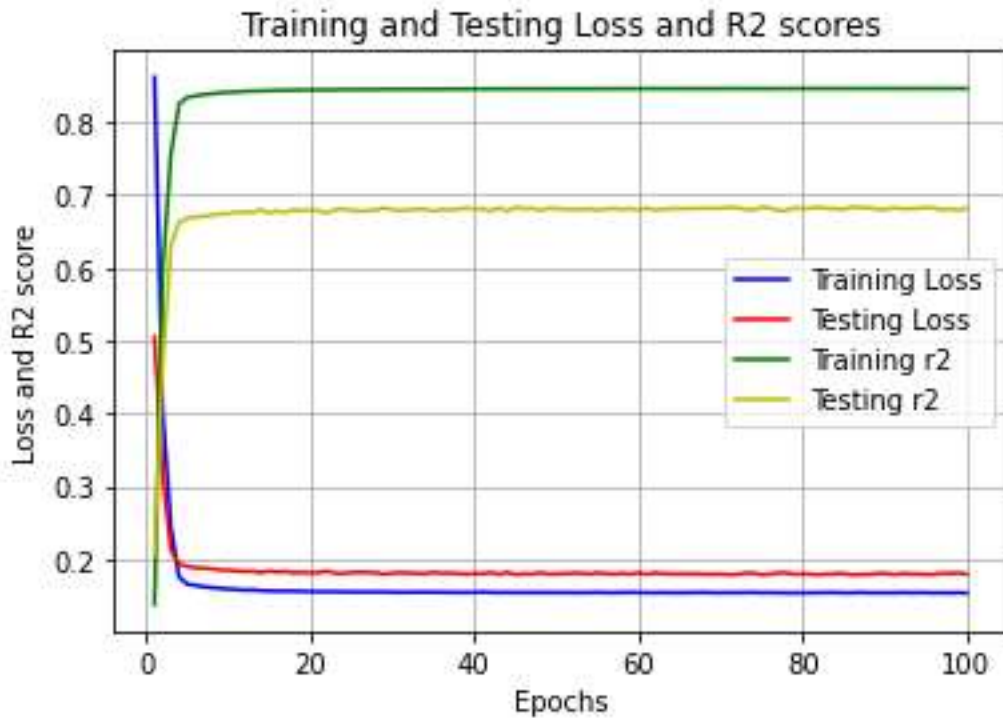


Figure 6-10 p1002 alternative sensors batch size 16 training and testing loss graph with r^2 scores.

In this instance, the training results alone do not provide the complete narrative, as demonstrated in Figure 6-11. Here, the orange prediction and blue test data against the calculated curve appear to fit better than the training scores would suggest. Consequently, Figure 6-12 delves deeper into this discrepancy, revealing that the neural network has successfully identified the curve within noisy data. This observation underscores the importance of not dismissing a model solely based on low test scores, as it may still yield valuable insights and accurate predictions.

The predictions generated by this simple neural network exhibit similarities to those produced by certain filtering algorithms utilized in control engineering. However, it's important to note that these filtering algorithms typically rely on first principles models rather than learned data from the system. While both approaches aim to achieve similar outcomes, their underlying methodologies and sources of information differ significantly.

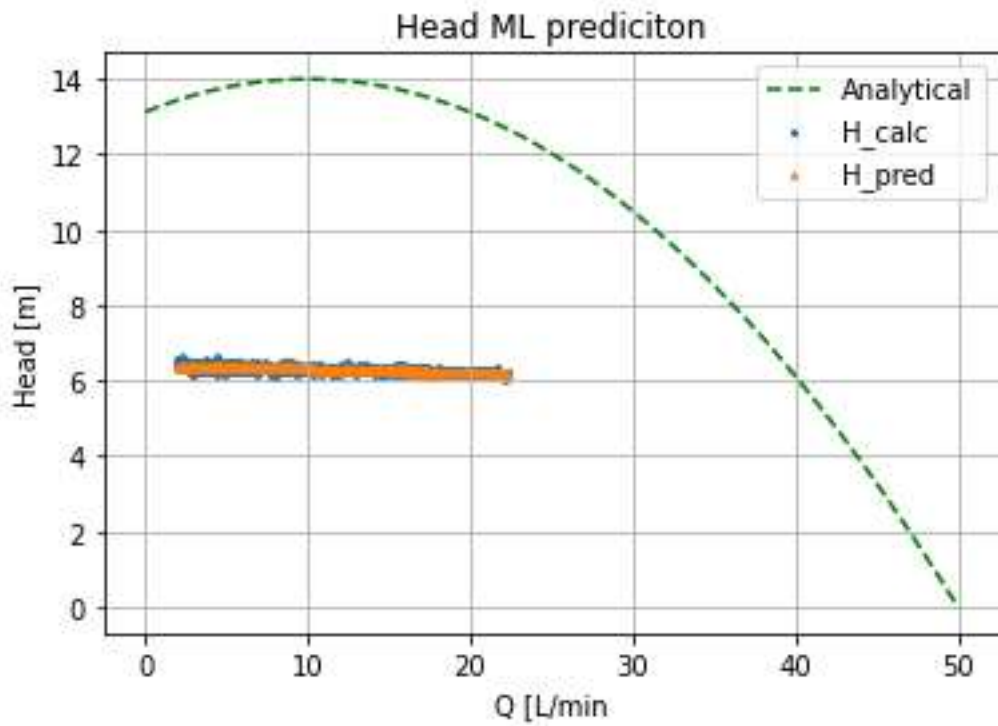


Figure 6-11 Original test data and predictions compared to analytical curve for P1002 alternative sensors.

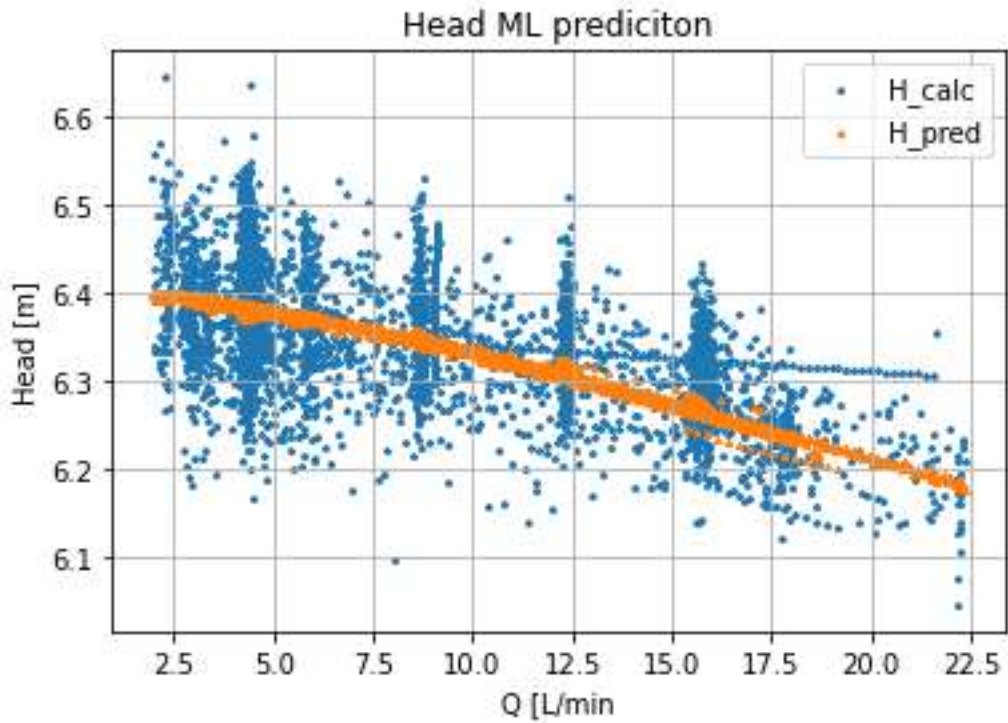


Figure 6-12 Original test data and predictions closer look for P1002 alternative sensors.

6.4 Additional tests

The results from additional tests conducted on P1001 using data from the alternative sensor suite are as follows:

6.4.1 Sigmoid versus tanh activation function

Testing the tanh activation function against the sigmoid activation function provided valuable insights, particularly regarding the rate of convergence. It was observed that tanh exhibited a notably faster convergence rate compared to sigmoid. Figure 6-13 depicts the plot of training a network with tanh over 100 epochs, while Figure 6-7, seen previously, serves as a reference for the same training with sigmoid.

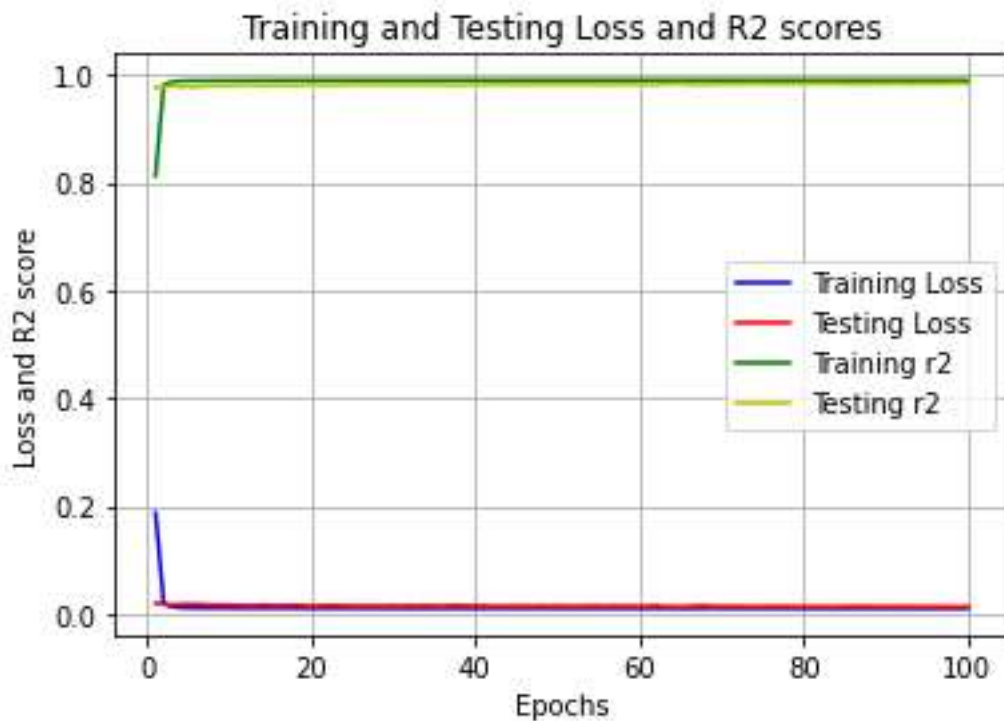


Figure 6-13 p1001 alternative training data with tanh activation function.

6 Machine learning results

Table 6-6 numerically presents the values, indicating that after 100 epochs, tanh leads by 0.05 percentage points.

Table 6-6 Test r^2 scores for tanh versus sigmoid activation function. Over select epochs up to 100

	Sigmoid	tanh
10	97.08%	97.91%
25	97.78%	98.04%
50	97.94%	98.28%
100	98.35%	98.40%

6.4.2 Including more data in the test set

Adjusting the test-train split to include the 75% operation point of the valve in the test dataset rather than the training data set for the p1001 network.

Comparing results for this test in Table 6-7 with the original in Table 6-4, shows a performance decrease of 2.21 percent points at the 100-epoch mark. The observed phenomenon may be attributed to the inherent complexity of the system and the amount of information contained within the specified range. Removing data points from the training set creates a substantial gap in the potential learning capacity of the model, leading to disruptions in its performance. This underscores the importance of comprehensive data coverage and the potential consequences of data selection on model robustness and effectiveness.

Table 6-7 changed test set for p1001 alternate sensor neural network scores. Over select epochs up to 100

Epoch	Batch 16 r^2 scores
10	95.16%
25	95.37%
50	95.45%
100	96.14%

6.4.3 Single output network results

Using the normal test-train split but modifying the network to have only one output and training it to estimate either head or watt resulted in the following outcomes, as shown in Figure 6-8, where Head estimation accuracy: 96.88% and Watt estimation accuracy: 99.68%

Interestingly, it's observed that the watt accuracy contributes significantly to the overall performance of the network, surpassing that of head measurements. This phenomenon can be attributed to the nature of a pump's H-Q curve, which is polynomial and hence more challenging for a linear model to estimate accurately. In contrast, watt calculations are linear in nature, facilitating more accurate estimation by the network.

Table 6-8 p1001 splitting the network in 2 r^2 scores for alternative sensor measurements batch size 16. Over select epochs up to 100

Epoch	r^2 scores head	r^2 scores watt
10	96.21%	99.24%
25	96.47%	99.32%
50	96.71%	99.51%
100	96.88%	99.68%

6.4.4 Adding features

Incorporating additional features and analyzing the results from training the network are detailed in Figure 6-9. Interestingly, the selected features in this instance led to a slight decrement in the network's performance, by 0.06 percentage points.

Table 6-9 R2 test scores for the network with one additional feature. Over select epochs up to 100

epochs	R^2 score with feature
10	97.78%
25	98.10%
50	98.25%
100	98.29%

6.5 Reducing the network down to matrix equations

This study explores the potential implementation of neural networks in lower-level hardware, considering the feasibility of utilizing matrix equations, scaling factors, biases, and activation functions such as the sigmoid function.

Table 6-10 and Table 6-11 present the weights and bias matrices for networks trained on datasets p1001 and p1002, respectively, using a batch size of 16 over 100 epochs. An intriguing aspect of this analysis is the observed differences in weight allocation to various inputs between the two networks, highlighting the distinct performance outcomes between them.

Table 6-10 weights and bias matrices for the alternative sensors p1001 after 100 epochs, shown layer by layer.

Layer n	Weights W_n	Biases b_n
1	$\begin{bmatrix} 0.0048 & -0.4653 & -0.5062 \\ 0.1023 & -1.1460 & 0.2461 \\ -0.0205 & -0.7216 & 1.0397 \\ 0.1004 & 0.6569 & 0.0674 \end{bmatrix}$	$\begin{bmatrix} -0.4375 \\ -0.8404 \\ 0.5502 \\ -0.2026 \end{bmatrix}$
2	$\begin{bmatrix} 0.9263 & 2.3149 & 0.2554 & -2.3008 \\ -3.8224 & 0.2062 & 2.8051 & 0.5724 \end{bmatrix}$	$\begin{bmatrix} -0.2489 \\ -0.4800 \end{bmatrix}$

Table 6-11 weights and bias matrices for the alternative sensors p1002 after 100 epochs, shown layer by layer.

Layer n	Weights W_n	Biases b_n
1	$\begin{bmatrix} 0.3926 & 1.2293 & -0.0282 \\ -0.0039 & -0.6268 & -0.2307 \\ 0.2074 & 0.0243 & -0.4030 \\ 0.2085 & -0.2636 & 0.4275 \end{bmatrix}$	$\begin{bmatrix} -0.7809 \\ -0.6503 \\ 0.0476 \\ 0.0889 \end{bmatrix}$
2	$\begin{bmatrix} -2.1342 & 1.6968 & 0.3633 & 1.1398 \\ 0.2739 & -1.7779 & -4.4569 & 0.4275 \end{bmatrix}$	$\begin{bmatrix} -0.6294 \\ 0.6663 \end{bmatrix}$

6.6 Transfer learning

The results from transfer learning tests.

6.6.1 Transfer learning test 1

The transfer learning results exceeded expectations, as depicted in Figure 6-14. This figure illustrates the baseline of the original neural network trained with the alternative sensors from P1001, serving as the model to build transfer learning from, with additional layers. Specific numerical values for select points in the figures are provided in Table 6-12.

6 Machine learning results

Additionally, Figure 6-15 shows that taking the original model and adding two new output nodes to it resulted in a performance surprisingly similar to training a model from scratch, as seen in Figure 6-10. This outcome is promising as it demonstrates the feasibility of transfer learning. With a better-performing model initially, further improvements in performance could be achieved.

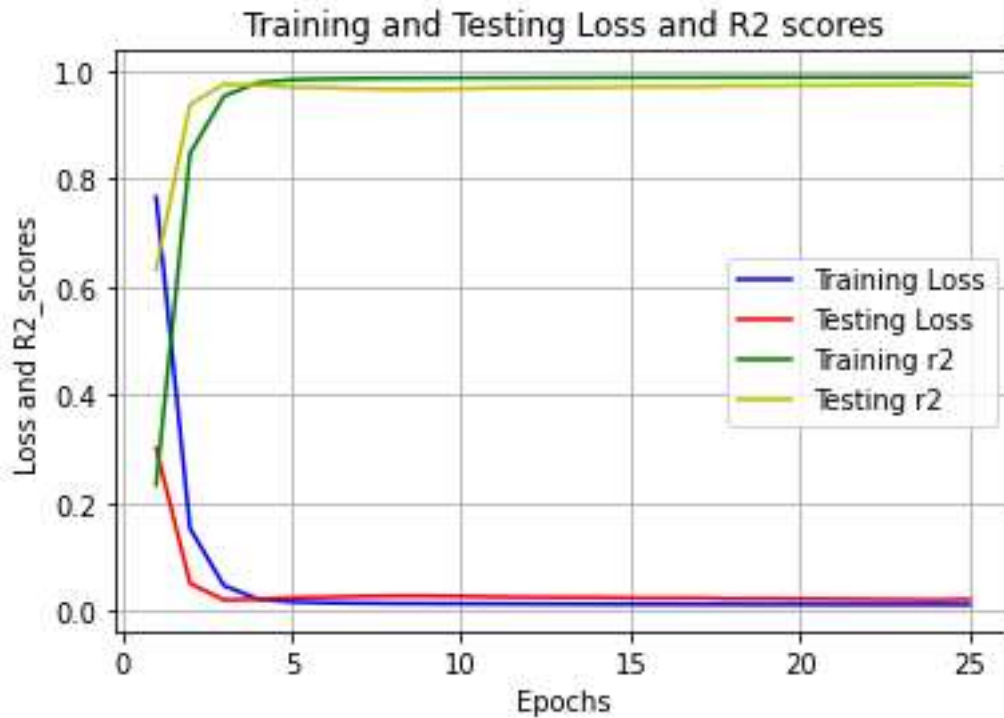


Figure 6-14 original model trained on data from p1001 with the alternate sensors.

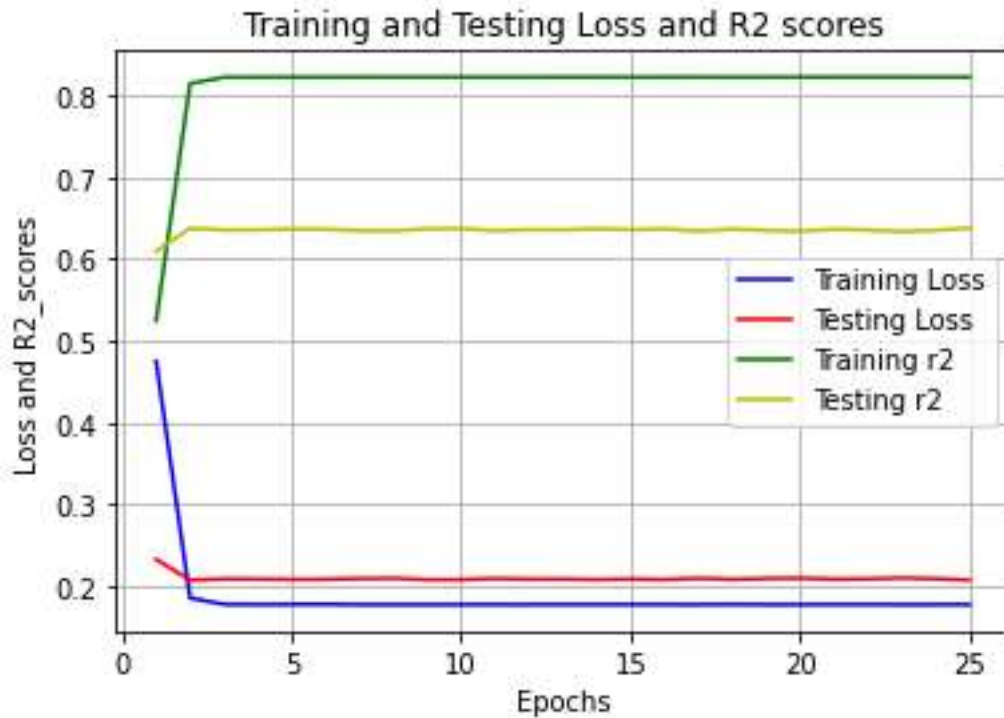


Figure 6-15 transfer learning model trained on data from p1002 using the original model from p1001, alternative sensor set.

Examining how the transfer learning model, initially trained on P1001 data, and then retrained on P1002 using transfer learning, predicts can provide valuable insights. Similar to previous subchapters, in Figure 6-16, the orange line represents the predicted values, the blue line represents the test data, and the green line represents the H-Q curve. Some overlap between the blue test data and the orange predicted values can be observed.

Figure 6-17 offers a clearer view, demonstrating how the model finds the curve within the noisy data. This suggests that the transfer learning approach effectively adapts the model to the new dataset, enabling it to capture the underlying patterns despite the presence of noise.

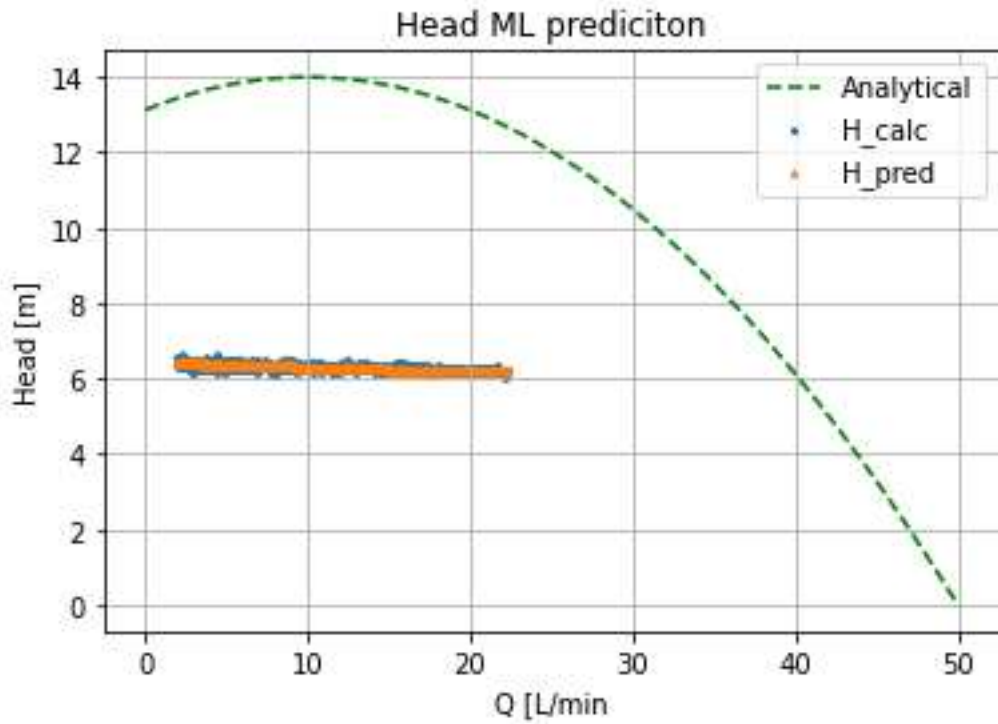


Figure 6-16 Original test data and predictions compared to analytical curve for P1002 alternative sensors using transfer learning model 1.

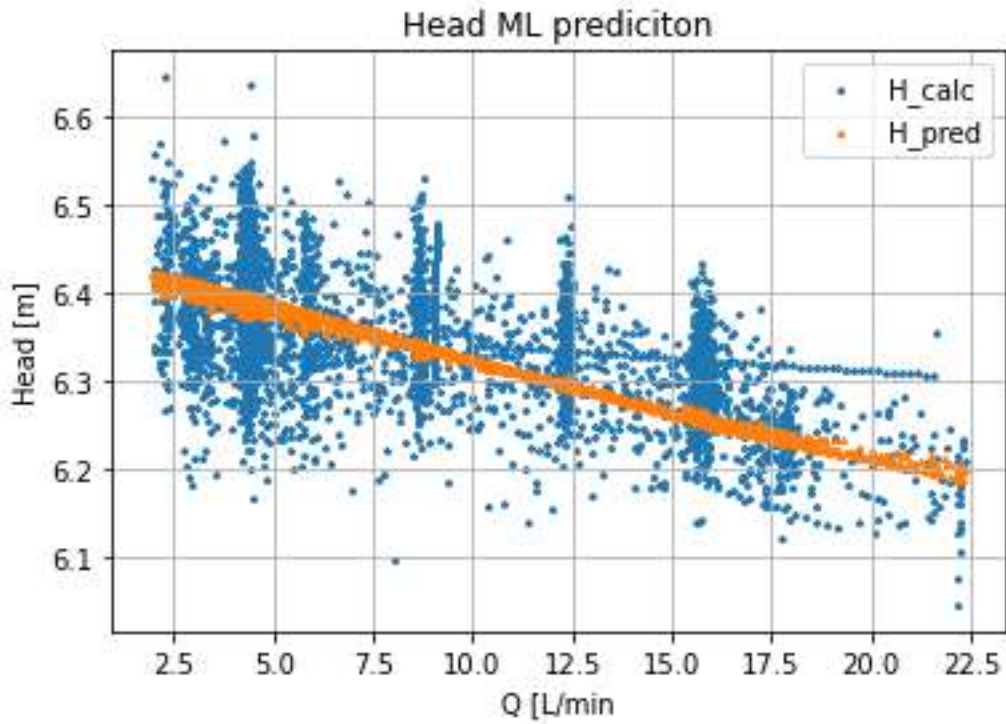


Figure 6-17 Original test data and predictions closer look for P1002 alternative sensors using transfer learning model 1.

6.6.2 Transfer learning test 2

In Figure 6-18, the same original model is utilized, but with the addition of a 3-node layer featuring a sigmoid function and the same 2 output nodes. Surprisingly, this configuration performs even better than the one from Figure 6-15, albeit only by approximately 1 percentage point.

Comparing the results from the transfer learning network with dedicated training for P1002 can also be observed in Table 6-12. At the 25-epochs mark, it shows approximately a 3-4% decrease in performance compared to the dedicated training for P1002. This difference could potentially be minimized further with a better base model and longer training, although this was not tested due to the unavailability of a superior base model.

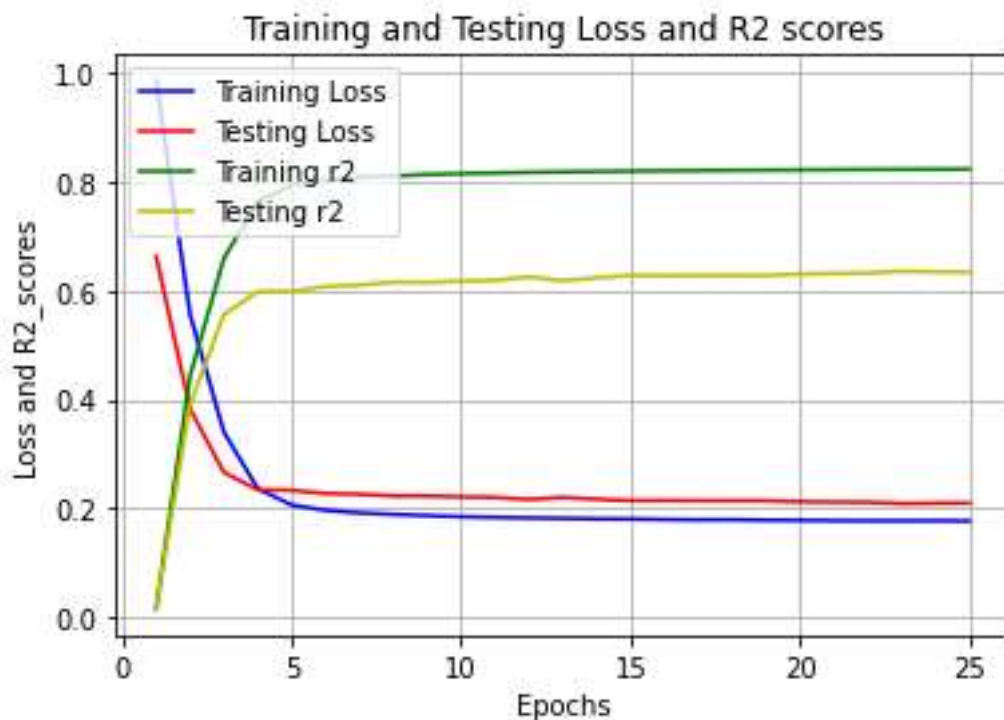


Figure 6-18 transfer learning model with 2 additional layers with a sigmoid activation function.

Figure 6-19 follows the convention of previous subchapters, where the orange line represents the predicted values, the blue line represents the raw data, and the green line represents the analytically sourced H-Q curve. Here, the transfer learning model demonstrates its capability to accurately identify the curve within the data.

Figure 6-20 offers a clearer visualization, illustrating how the predictions curve aligns with the noisy data. This further emphasizes the effectiveness of the transfer learning approach in capturing the underlying patterns within the dataset.

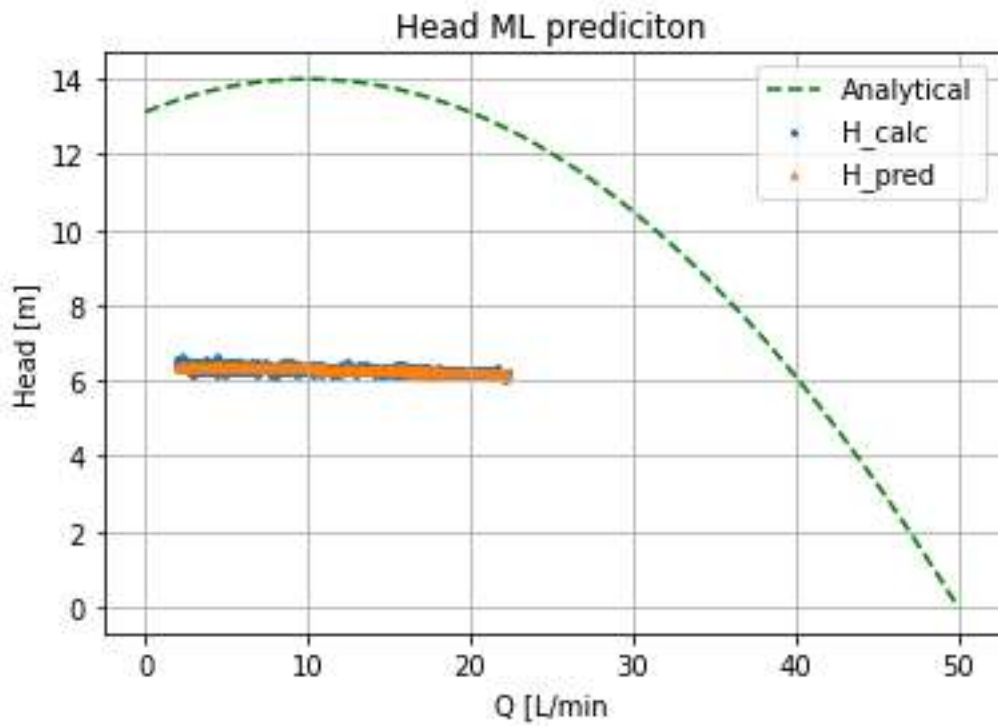


Figure 6-19 Original test data and predictions closer look for P1002 alternative sensors using transfer learning model 2.

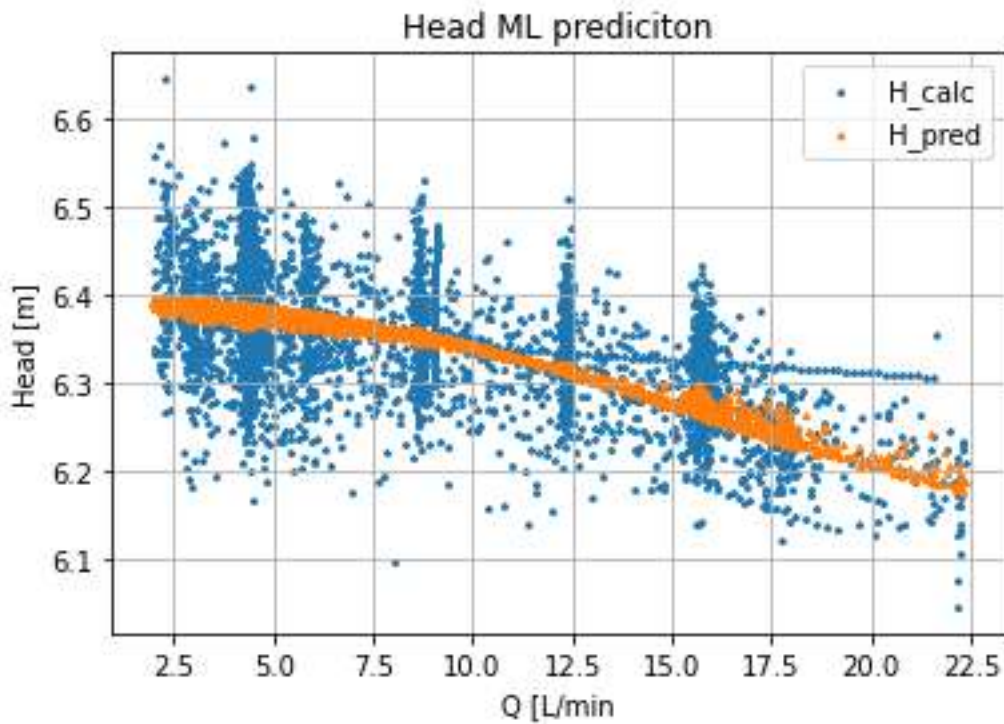


Figure 6-20 Original test data and predictions closer look for P1002 alternative sensors using transfer learning model 2.

6 Machine learning results

Table 6-12 R² test scores for the different transfer learning tests over 25 epochs. Over select epochs up to 25

Epochs	R ² test scores original model, p1001	R ² test scores one additional layer transfer model, p1002	R ² test scores for two additional layers with activation function, p1002	Dedicated training p1002
5	96.31%	63.55%	61.04%	66.53%
10	96.73%	63.58%	62.98%	67.04%
15	97.13%	63.77%	64.09%	67.66%
25	97.55%	63.65%	64.36%	67.47%

7 Discussion

This chapter is for comparing and discussing results versus other results versus the theory/method behind it.

7.1 Interpolation versus extrapolation

This discussion addresses the rationale behind showcasing the full range of the pumps from 0 to 50 L/min, despite the data being confined to the 0 to 20 L/min range.

Figure 5-5 and Figure 5-6 both extrapolate data to accommodate the pump's rated flow. While this extrapolation provides a useful visualization, it may introduce inaccuracies by depicting data beyond the experimental range.

In contrast, when focusing on interpolation and examining point-to-point data, it becomes evident that no curve fit line can accurately represent all data points, particularly in the presence of random noise processes during data collection. This discrepancy is apparent when comparing the area covered by the raw data to that covered by the curve fit line on the graphs.

7.2 Random forest regressor versus neural networks

Comparing the random forest regressor to a neural network is a valuable exercise, considering factors such as simplicity, ease of implementation, accuracy, efficiency, and compatibility with control systems.

While the random forest regressor is simpler and easier to implement in code, the comparison should also account for factors such as efficiency, accuracy, and data flow within the system. When weighing these considerations, it becomes apparent that a neural network may be the preferred choice in this case. Despite its potentially higher complexity, a neural network offers superior performance and versatility, making it well-suited for prediction and implementation into a control system.

7.3 Correct versus alternative sensors in machine learning

Comparing the performance of neural networks trained using alternative sensor suits versus the normal sensor suits discussed in Chapter 4.4 reveals significant differences in results.

Taking P1001 as an example, with the normal sensor suit, the network's predictions demonstrated high coverage and accuracy, leading to potential overfitting. However, it failed to capture the pump's curve within the data. In contrast, utilizing the alternative sensor suit resulted in predictions that closely resembled the pump's H-Q curve, despite reflecting all the noise present in the data. This suggests that the network's simplicity led to outcomes resembling more of a filtering algorithm rather than a curve-fitting approach.

Similar observations apply to P1002, where the higher degree of noise in the head calculations further emphasized the network's ability to identify the curve within the data, even with seemingly lower accuracy.

These findings imply that gathering data with normal sensors and training using alternative sensors can yield a robust representation of the pump's curve using a simple neural network architecture.

This refers to Figure 6-3 and Figure 6-9, where we see that the regular sensors match the training data closely, while the alternative sensors show a more distinct H-Q curve pattern from the data.

7.4 Usefulness in condition monitoring

Condition monitoring plays a critical role in ensuring the optimal performance and longevity of pumps. While various methods were explored in this study, rigorous testing was not conducted, leading to their inclusion in the discussion chapter rather than the method and results chapters. However, potential methods for condition monitoring will be outlined here, leveraging the understanding gained from estimating the pump head and H-Q curve.

The H-Q curve serves as a fundamental parameter in pump operation and design selection for a given system. However, changes over time, including wear and tear on the pump, pipes, and fittings, alterations to the process, and external factors, can impact the pump's performance.

Effective condition monitoring requires a comprehensive dataset, including measurements for head, vibrations, temperature, flow, and power, among others. By analyzing the combination of these parameters, valuable insights into the pump's condition can be obtained.

When only flow out and tank level measurements are available, interpreting them as pressure in and amperage may provide limited information about the pump. Therefore, estimating the head and the pump's H-Q curve can offer valuable insights into its condition. Monitoring changes in the head estimate over time can serve as an early indicator of potential issues, alerting operators or maintenance personnel to take corrective action.

Understanding how the pump is operated and where it operates on the pump curve is crucial for its longevity. Below are some examples of how condition monitoring could be performed.

7.4.1 Rule based monitoring

Rule-based monitoring involves using simple rules or criteria, rather than sophisticated intelligent systems, to analyze data and make decisions. In this context, operators rely on estimated plots and basic insights to optimize pump operations.

One approach to rule-based monitoring is to utilize the flow sensor output from the pump in conjunction with an analytical H-Q curve to gain insight into the pump's performance relative to its breakeven point. The breakeven point typically lies in the middle of the specified flow range, with margins on either side.

By associating each flow with an estimated head, upper and lower limits can be established. These limits represent acceptable ranges within which the pump should operate. Alarms can then be configured to alert operators when the pump exceeds these limits, indicating that it is operating outside its optimal performance range.

This approach allows operators to monitor pump performance in real-time and take corrective actions promptly to ensure efficient and reliable operation. While it may not involve sophisticated data analysis techniques, rule-based monitoring can still be effective in identifying and addressing operational issues in a timely manner.

7.4.2 Smart monitoring

Smart monitoring builds upon the foundation of rule-based monitoring by incorporating more automated logic and artificial intelligence (AI) capabilities. In contrast to using the analytical H-Q curve, smart monitoring relies on experimental curves specific to the system in which the pump operates.

Rather than relying solely on predetermined limits based on analytical curves, smart monitoring dynamically adjusts limits based on real-time data and historical performance. By continuously analyzing how the pump operates over time within its specific system, smart monitoring can adapt and optimize limits to reflect the pump's actual operating conditions and performance characteristics.

This approach enables more proactive and adaptive monitoring, allowing for early detection of deviations from expected behavior and prompt intervention to prevent issues before they escalate. By leveraging AI algorithms and automated logic, smart monitoring enhances operational efficiency, reliability, and overall pump performance.

7.4.3 Examples of rule-based monitoring implementation

Figure 7-1 shows the potential limits used on p1001 with the flow plotted on the analytical curve. In this case all measured flows fall under the low limit.

Figure 7-2 shows the potential limits used on p1002 with the flow plotted on the analytical curve, in this case only some measured flows fall inside the correct operating range.

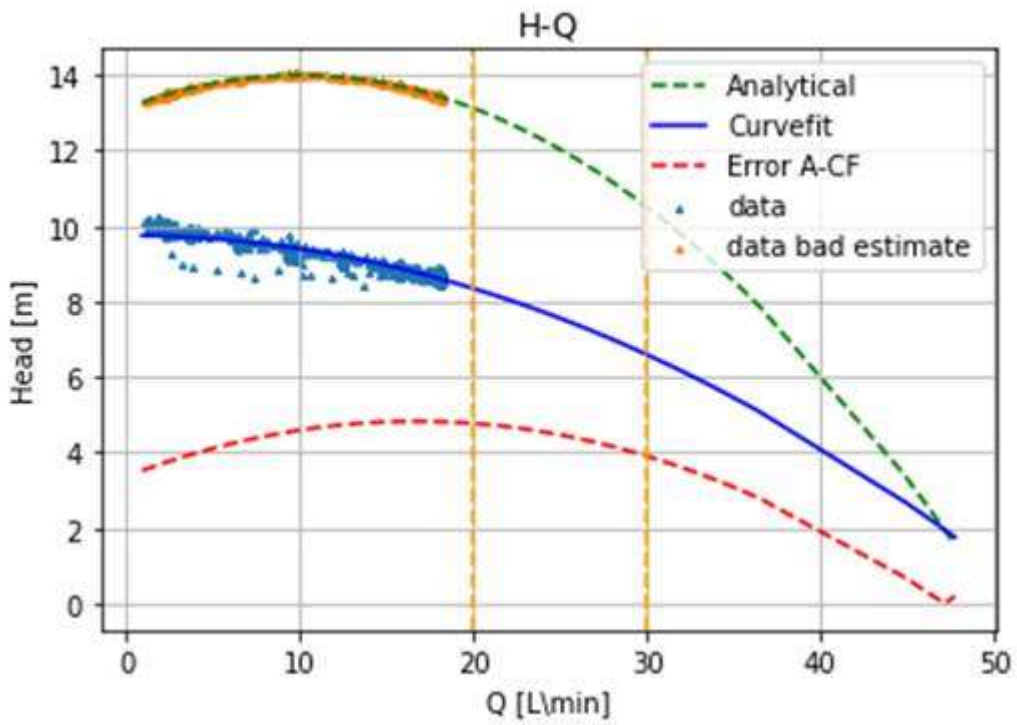


Figure 7-1 p1001 data plot with high and low limits in orange at 30 and 20 L/min respectively.

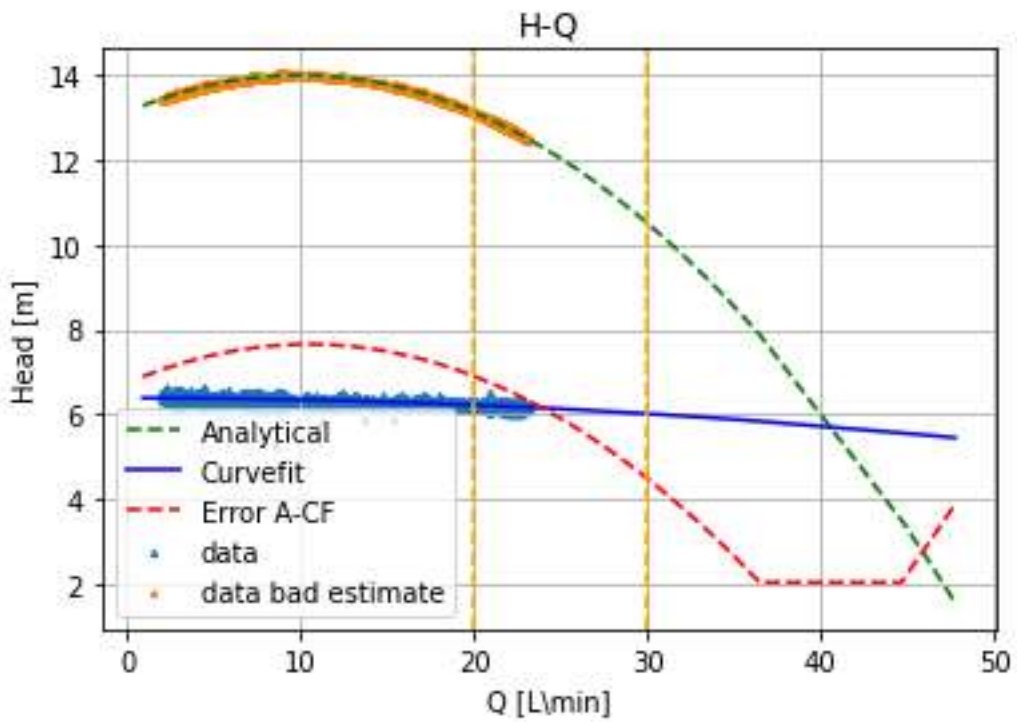


Figure 7-2 p1002 data plot with high and low limits in orange at 30 and 20 L/min respectively.

8 Conclusion

In conclusion, this thesis has outlined a comprehensive approach to understanding and utilizing pump system dynamics through a combination of experimental design and machine learning techniques. By leveraging sensor data and accurate calculations, we can effectively determine pump head, while curve fitting methods enable the characterization of system-specific H-Q curves. Furthermore, machine learning algorithms, when trained on high-quality datasets, offer the capability to estimate H-Q curves using alternative sensor data, providing flexibility and scalability in pump system analysis.

Moreover, transfer learning has been explored to transfer knowledge from one pump system to another, demonstrating potential but also highlighting the importance of continuous improvement in base models for optimal performance. Overall, this thesis underscores the significance of data quality, appropriate methodology, and ongoing refinement in achieving accurate and reliable predictions in pump system analysis and prediction.

8.1 Future work

Designing an experimental setup that would yield close to zero deviation between the analytical curve and the curve fit curve is theoretically possible but would require careful consideration of several factors. These factors include precise control over experimental conditions, accurate measurement devices, and thorough understanding of the system dynamics. Achieving such a setup would involve extensive calibration and validation processes to ensure accuracy and reliability.

Regarding the transfer learning aspect, transferring a model trained on data from an optimal setup to a less optimal setup can still yield valuable insights, although performance may be impacted. The model may need to undergo further fine-tuning or adaptation to accommodate the differences in the new setup. However, having a close-to-perfect model as a starting point can provide a strong foundation for transfer learning, potentially enabling the model to adapt more effectively to the new conditions.

Using the close-to-perfect model to estimate conditions and then feeding those estimates into another network for further analysis is a plausible approach. The effectiveness of this approach would depend on various factors, including the accuracy of the initial model, the complexity of the conditions being estimated, and the capabilities of the subsequent network. Overall, leveraging a high-quality model for estimation purposes can enhance the overall performance and reliability of the system.

References

- [1] IWAKI, “MX Brochure UK,” Available: https://iwaki-nordic.com/literature/process_pumps/mag_drive_plastic/mx/brochure/MX_brochure_uk.pdf (accessed may.8, 2024)
- [2] E. Salomonsen, U. Shamir M. Housh, “Optimization Methodology for Estimating Pump Curves Using SCADA Data”, *Water* **2021**, 13, 586. <https://www.mdpi.com/2073-4441/13/5/586> (accessed may.8, 2024)
- [3] L. Bachus and A. Custodio, «Understanding Pump Curves,» in *Know and Understand Centrifugal Pumps*,., 1, Oxford, UK; Elsevier, 2003, chapter 7, 76.
- [4] Carsten Ramberg, Borregaard, 12.0.2024
- [5] E. Hoffer, I. Hubara, and D. Soudry, “Train longer, generalize better: closing the generalization gap in large batch training of neural networks,” *arXiv [stat.ML]*, May 24, 2017. [Online]. Available: <http://arxiv.org/abs/1705.08741>
- [6] A. J. Wheeler, A. R. Ganji, “Guidelines for Planning and Documenting Experiments” in *Introduction to Engineering Experimentation*, Pearson/Prentice Hall, 2004, 12,
- [7] M. Ceraolo and D. Poli, “6.4 power in three-phase systems,” in *Fundamentals of electric power engineering: From electromagnetics to power systems*, 1st ed., 2014. [Online]. Available: <http://ndl.ethernet.edu.et/bitstream/123456789/30644/1/31.pdf>
- [8] A. J. Wheeler, A. R. Ganji, “Statistical Analysis in experimental data” in *Introduction to Engineering Experimentation*, Pearson/Prentice Hall, 2004, 6,
- [9] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems*, 2nd ed. O’Reilly Media, Inc., 2019.
- [10] “PyTorch,” www.pytorch.org. <https://pytorch.org/get-started/locally/> (accessed may.8, 2024)
- [11] S. P. Berge, B. F. Lund, and R. Ugarelli, “Condition Monitoring for Early Failure Detection. Frognerparken Pumping Station as Case Study,” *Procedia Engineering*, vol. 70, pp. 162–171, 2014, doi: <https://doi.org/10.1016/j.proeng.2014.02.019>.
- [12] Turunen, T., Miettinen, J., Hämäläinen, A., Karhinen, A., Viitala, R. (2023). Deep Learning for Centrifugal Pump Condition Monitoring Using Data from Variable Frequency Drive. In: Okada, M. (eds) *Advances in Mechanism and Machine Science*. IFToMM WC 2023. *Mechanisms and Machine Science*, vol 147. Springer, Cham. https://doi-org.ezproxy1.usn.no/10.1007/978-3-031-45705-0_88t
- [13] OpenAI, “ChatGPT,” OpenAI, 2023. <https://openai.com/chatgpt> (accessed may.8, 2024)
- [14] NumPy, “Overview — NumPy v1.19 Manual,” numpy.org, 2022. <https://numpy.org/doc/stable/> (accessed may.8, 2024)
- [15] Pandas, “pandas documentation — pandas 1.0.1 documentation,” pandas.pydata.org. <https://pandas.pydata.org/docs/> (accessed may.8, 2024)

References

- [16] “SciPy documentation — SciPy v1.8.1 Manual,” docs.scipy.org.
<https://docs.scipy.org/doc/scipy/> (accessed may.8, 2024)
- [17] Matplotlib, “Matplotlib: Python plotting — Matplotlib 3.1.1 documentation,”
Matplotlib.org, 2012. <https://matplotlib.org/> (accessed may.8, 2024)
- [18] Scikit-learn, “scikit-learn: machine learning in Python,” Scikit-learn.org, 2019.
<https://scikit-learn.org/stable/> (accessed may.8, 2024)
- [19] E. Raff, Inside deep learning: Math, Algorithms, Models. Simon and Schuster, 2022. 1,
Ch. 13
- [20] N. Seth, “Estimation of neurons and forward propagation in neural net,” Analytics
Vidhya, Jul. 20, 2023. [https://www.analyticsvidhya.com/blog/2021/04/estimation-of-
neurons-and-forward-propagation-in-neural-net/](https://www.analyticsvidhya.com/blog/2021/04/estimation-of-neurons-and-forward-propagation-in-neural-net/) (accessed may.8, 2024)
- [21] Japan, Eu, China, and T. Mx, “IWAKI MAGNETIC DRIVE PUMPS Patent.” Available:
https://iwaki-nordic.com/literature/process_pumps/mag_drive_plastic/mx/broc (accessed
may.8, 2024)

Appendices

Appendix A Thesis task description

FMH606 Master's Thesis

Title: Data-driven Approaches for Pump Condition Monitoring and Curves Estimation

USN supervisor: Ru Yan, Saba Mylvaganam

External partner: Dag Harald Skjeltorp/ BORREGAARD, Martin Forsberg Lie/ BORREGAARD

Task background:

Borregaard, <https://www.borregaard.com/>, with its extensive chemical production facilities, specializing in the manufacture of cellulose and lignin, has an extensive network of pumps ranging from 1000 to 3000 units in daily operation. It is challenging to monitor the performance of all these pumps using the associated sensor data to ascertain if all these pumps are operating according to their respective performance specifications. Notably, many of these pumps handle not only water, but also several types of hazardous chemicals. Addressing a pump failure necessitates immediate attention by maintenance engineers. Preventive maintenance can alleviate costly and hazardous repairs and replacement of defective pumps requiring strict safety measures.

Borregaard is conducting a pilot program to test and evaluate the feasibility of combining already installed (or being installed) sensors to monitor pumps' performance and extend their operational lifespans using data-driven based sensor data fusion for enabling predictive maintenance.

Task description:

1. List common causes of failures and replacements of industrial pumps.
2. Description of a typical network of pumps from a selected section of the flow loop with the necessary P&ID, with information all the tags available and the details of measurands and their ranges in the plant and the corresponding measurement ranges of the sensors/instruments in the P&ID.
3. Conduct a literature survey on soft sensor methods for pump condition monitoring and estimation of pump curves.
4. Assemble a dataset by running the test rig based on a pre-defined experimental design and collect relevant data.
5. Conduct comprehensive data analysis including development and validation of data-driven models for continuously monitoring pump performance and generating updated pump curves.
6. Refine the experimental design to determine the test rig configurations needed for optimal data collection.
7. Evaluate the potential performance degradation when replacing traditional hard sensors with data-driven based soft sensors.
8. If task 7 leads to promising results, increase the number and types of pumps and test the data-driven models with them.
9. (Optional) Explore the potential for extending the model to enable predictive maintenance by identifying pumps showing signs of failure.
10. (Optional) Investigate the feasibility of integrating the model in a fault detection system for estimate remaining useful life (RUL) of pumps with enhanced data-driven models leading to simpler rule-based methods useful for maintenance engineers.

Student category: IIA

Is the task suitable for online students (not present at the campus)? Reserved for Kristian Sande Sjølyst.

Practical arrangements:

Hardware and software will be available from Borregaard with necessary support.

Supervision:

As a general rule, the student is entitled to 15-20 hours of supervision. This includes necessary time for the supervisor to prepare for supervision meetings (reading material to be discussed, etc).

Signatures:

Supervisor (date and signature):

Ru Yan Digitally signed by Ru Yan
Date: 2024.02.01 13:41:54
+01'00'

Student (write clearly in all capitalized letters): *Kristian Sande Sjølyst*

Student (date and signature): *01/01/2024 Kristian Sande Sjølyst*

Appendix B GitHub repository archive:

[n3cromans3r/FMH606_162562: code developed for masters thesis \(github.com\)](https://github.com/n3cromans3r/FMH606_162562)